Lesson 4
Organizing Epidemiologic Data

When we collect more records than we can review individually, we can use tables, graphs, and charts to organize, summarize, and display the data clearly and effectively. With tables, graphs, and charts we can analyze data sets of a few dozen or a few million. These tools allow us to identify, explore, understand, and present distributions, trends, and relationships in the data. Thus tables, graphs, and charts are critical tools not only when we perform descriptive and analytic epidemiology, but also when we need to communicate our epidemiologic findings to others.

Objectives

After studying this lesson and answering the questions in the exercises, a student will be able to do the following:

- Correctly prepare tables with one, two, or three variables
- Correctly prepare the following types of graphs: arithmetic-scale line graphs, semilogarithmic-scale line graphs, histograms, frequency polygons, and scatter diagrams
- Correctly prepare the following types of charts: bar charts, pie charts, spot maps, area maps, and box plots
- Describe when to use each type of table, graph, and chart

Introduction to Tables, Graphs, and Charts

Data analysis is an important component of epidemiologic practice. To analyze data effectively, an epidemiologist must first become familiar with the data before applying analytic techniques. The epidemiologist may begin by examining individual records such as those contained in a line listing, but will quickly progress to summarizing the data with tables. Sometimes, the resulting tables are the only analysis that is needed, particularly when the amount of data is small and relationships are straightforward. When the data are more complex, graphs and charts can help the epidemiologist visualize broader patterns and trends and identify variations from those trends. Variations may represent important new findings or only errors in typing or coding which need to be corrected. Thus, tables, graphs, and charts are essential to the verification and analysis of the data.

Once an analysis is complete, tables, graphs, and charts further serve as useful visual aids for describing the data to others. In preparing tables, graphs, and charts for others, you must keep in mind that their primary purpose is to communicate information about the data.

Tables

A table is a set of data arranged in rows and columns. Almost any quantitative information can be organized into a table. Tables are useful for demonstrating patterns, exceptions, differences, and other relationships. In addition, tables usually serve as the basis for preparing more visual displays of data, such as graphs and charts, where some of the detail may be lost.

Tables designed to present data to others should be as simple as possible. Two or three small tables, each focusing on a different aspect of the data, are easier to understand than a single large table that contains many details or variables.

A table should be self-explanatory. If a table is taken out of its original context, it should still convey all the information necessary for the reader to understand the data. To create a table that is self-explanatory, follow the guidelines below:

- Use a clear and concise title that describes the what, where, and when of the data in the table. Precede the title with a table number (for example, Table 4.1).
- Label each row and each column clearly and concisely and include the units of measurement for the data (for example, years, mm Hg, mg/dl, rate per 100,000).
- Show totals for rows and columns. If you show percents (%), also give their total (always 100).
- Explain any codes, abbreviations, or symbols in a footnote. (for example, *Syphilis P&S = primary and secondary syphilis*)
- Note any exclusions in a footnote (*1 case and 2 controls with unknown family history were excluded from this analysis*).
- Note the source of the data in a footnote if the data are not original.

One-Variable Table

In descriptive epidemiology, the most basic table is a simple frequency distribution with only one variable, such as Table 4.1a. (Frequency distributions are discussed in Lessons 2 and 3.) In such a frequency distribution table, the first column shows the values or categories of the variable represented by the data, such as age or sex. The second column shows the number of persons or events that fall into each category.

Often, a third column lists the percentage of persons or events in each category, as in Table 4.1b. Note that the percentages in Table 4.1b add up to 100.1% rather than 100.0% due to rounding to one decimal place. This is commonly true in tables that show percentages. Nonetheless, the total percent should be given as 100.0%, and a footnote explaining that the difference is due to rounding should be included.

Table 4.1a
Primary and secondary syphilis morbidiy
by age, United States, 1989

Age group (years)	Number of cases
≤14	230
15-19	4,378
20-24	10,405
25-29	9,610
30-34	8,648
35-44	6,901
45-54	2,631
≥55	1,278
Total	**44,081**

Source: 12

Table 4.1b
Primary and secondary syphilis morbidity
by age, United States, 1989

Age group	Cases	
(years)	Number	Percent
≤14	230	0.5
15-19	4,378	10.0
20-24	10,405	23.6
25-29	9,610	21.8
30-34	8,648	19.6
35-44	6,901	15.7
45-54	2,631	6.0
≥55	1,278	2.9
Total	**44,081**	**100.0***

*Percentages do not add to 100.0% due to rounding.
Source: 12

The one-variable table can be further modified to show either cumulative frequency or cumulative percent, as in Table 4.1c. We now see that 75.5% of the primary and secondary syphilis cases occurred in persons less than 35 years old.

Table 4.1c
Primary and secondary syphilis morbidity by age, United States, 1989

Age group (years)	Cases		
	Number	Percent	Cumulative %
≤14	230	0.5	0.5
15-19	4,378	10.0	10.5
20-24	10,405	23.6	34.1
25-29	9,610	21.8	55.9
30-34	8,648	19.6	75.5
35-44	6,901	15.7	91.2
45-54	2,631	6.0	97.2
≥55	1,278	2.9	100.0
Total	**44,081**	**100.0***	**100.0%**

*Percentages do not add to 100.0% due to rounding.
Source: 12

Two- and Three-Variable Tables

Tables 4.1a, 4.1b, and 4.1c show case counts (frequency) by only one variable: age. Data can also be cross-tabulated to show counts by a second variable. Table 4.2 shows the number of syphilis cases by both age and sex of the patient.

A two-variable table with cross-tabulated data is also known as a **contingency table**. Table 4.3 is an example of a common type of contingency table, which is called a **two-by-two table** because each of the two variables has two categories. Epidemiologists frequently use contingency tables to display the data used in calculating measures of association and tests of statistical significance.

Table 4.2
Newly reported cases of primary and secondary syphilis by age and sex, United States, 1989

Age group (years)	Number of cases by sex		
	Male	Female	Total
≤14	40	190	230
15-19	1,710	2,668	4,378
20-24	5,120	5,285	10,405
25-29	5,304	4,306	9,610
30-34	5,537	3,111	8,648
35-44	5,004	1,897	6,901
45-54	2,144	487	2,631
≥55	1,147	131	1,278
Total	**26,006**	**18,075**	**44,081**

Source: 12

Epidemiologists also use two-by-two tables to study the association between an exposure and disease. These tables are convenient for comparing persons with and without the exposure, and those with and without the disease. Table 4.4 shows the generic format of such a table. As shown there, disease status (e.g., ill versus well) is usually designated along the top of the table, and exposure status (e.g., exposed versus not exposed) is designated along the side. The letters a, b, c, and d within the 4 cells of the two-by-two table refer to the number of persons with the disease status indicated above and the exposure status indicated to its left. For example, in Table 4.4, **c** is the number of persons in the study who have the disease, but who did not have the exposure being studied. Note that the **"H"** in the row totals **H1** and **H2** stands for horizontal; the **"V"** in the column total **V1** and **V2** stands for vertical. The total number of subjects included in the two-by-two table is represented by the letter **T** (or **N**).

When displaying data to others, it is best to use one- or two-variable tables, like those on the preceding pages. Sometimes, however, you may want to include a third variable to show a set of data more completely. Table 4.5 shows such a three-variable table for the variables of age, race, and sex. As you can see, a three-variable table is rather busy. It is the maximum amount of complexity you should ever include in a single table.

Table 4.3
Follow-up status among diabetic and nondiabetic white men
NHANES follow-up study, 1982-1984

	Dead	Alive	Total	Percent dead
Diabetic	100	89	189	52.9
Nondiabetic	811	2,340	3,151	25.7
Total	**911**	**2,429**	**3,340**	

Source: 18

Table 4.4
General format for 2 x 2 table

	Ill	Well	Total
Exposed	a	b	**H1**
Unexposed	c	d	**H2**
Total	**V1**	**V2**	**T**

Table 4.5
Primary and secondary syphilis morbidity by age, race, and sex, United States, 1989

		Race			
Age (years)	**Sex**	**White**	**Black**	**Other**	**Total**
≤14	Male	2	31	7	40
	Female	14	165	11	190
	Total	16	196	18	230
15-19	Male	88	1,412	210	1,710
	Female	253	2,257	158	2,668
	Total	341	3,669	368	4,378
20-24	Male	407	4,059	654	5,120
	Female	475	4,503	307	5,285
	Total	882	8,562	961	10,405
25-29	Male	550	4,121	633	5,304
	Female	433	3,590	283	4,306
	Total	983	7,711	916	9,610
30-34	Male	564	4,456	520	5,537
	Female	316	2,628	167	3,111
	Total	880	7,081	687	8,648
35-44	Male	654	3,858	492	5,004
	Female	243	1,505	149	1,897
	Total	897	5,363	641	6,901
45-54	Male	323	1,619	202	2,144
	Female	55	392	40	487
	Total	378	2,011	242	2,631
≥55	Male	216	823	108	1,147
	Female	24	92	15	131
	Total	240	915	123	1,278
Total for all ages	Male	2,804	20,376	2,826	26,006
	Female	1,813	15,132	1,130	18,075
	Total	**4,617**	**35,508**	**3,956**	**44,081**

Source: 12

Exercise 4.1

The data in Table 4.6 describe characteristics of the 36 residents of a nursing home during an outbreak of diarrheal disease.

A. Construct a table of the illness (diarrhea) by menu type. Use diarrhea status as column labels and menu types as row labels.

B. Construct a two-by-two table of the illness (diarrhea) by exposure to menu A.

Answers on page 269.

Table 4.6
Characteristics of residents of Nursing Home A
during outbreak of diarrheal disease, January, 1989

Resident no.	Age	Sex	Room	Menu	Diarrhea?	Date of onset
1	71	F	103	A	Yes	1/15
2	72	F	105	A	Yes	1/23
3	74	F	105	A	No	
4	86	F	107	B	No	
5	83	F	107	B	No	
6	68	F	109	A	Yes	1/18
7	69	F	109	C	No	
8	64	F	111	A	Yes	1/16
9	66	M	111	A	Yes	1/18
10	68	M	104	A	Yes	1/20
11	70	M	106	A	No	
12	86	M	110	A	No	
13	73	M	112	B	No	
14	82	M	219	C	No	
15	72	M	221	C	No	
16	70	M	221	B	No	
17	77	M	227	D	No	
18	80	M	227	D	No	
19	71	F	231	A	Yes	1/14
20	68	F	231	D	Yes	1/15
21	64	F	233	A	No	
22	73	F	235	A	Yes	1/13
23	75	F	235	B	No	
24	78	F	222	C	No	
25	72	F	222	A	No	
26	66	M	224	B	No	
27	69	M	226	A	Yes	1/16
28	75	M	228	E	No	
29	71	M	230	A	Yes	1/13
30	83	M	232	F	No	
31	84	M	232	D	No	
32	79	M	234	A	Yes	1/12
33	72	M	234	D	Yes	1/14
34	77	M	236	A	Yes	1/13
35	78	M	236	B	No	
36	80	M	238	D	No	

Tables of Other Statistical Measures

Tables 4.1 through 4.3 show case counts (frequency). The cells of a table can just as easily contain means, rates, years of potential life lost, relative risks, and other statistical measures. As with any table, the title and headings must clearly identify what data are presented. For example, both the title and the top heading of Table 4.7 indicate that rates are presented.

Table 4.7
Newly reported cases of primary and secondary syphilis,
age- and race-specific rates per 100,000 (civilian) population
United States, 1989

Age group (years)	Rate (per 100,000) by race			
	White	Black	Other	Total
≤14	0.0	2.4	0.8	0.4
15-19	2.4	131.5	51.0	24.3
20-24	5.8	323.0	139.2	55.9
25-29	5.4	270.9	117.9	44.1
30-34	4.7	256.6	83.2	38.8
35-44	2.9	135.0	47.8	19.0
45-54	1.7	76.7	29.6	10.5
≥55	0.5	19.4	10.4	2.4
Total	**2.2**	**115.8**	**45.8**	**17.7**

Source: 12

Table Shells

Although we cannot analyze data before we have collected them, we should design our analyses in advance to expedite the analysis once the data are collected. In fact, most protocols, which are written before a study can be conducted, require a description of how the data will be analyzed. As part of the analysis plan, we develop **table shells** which show how the data will be organized and displayed. Table shells are tables that are complete except for the data. They show titles, headings, and categories. In developing table shells that include continuous variables such as age, we create more categories than we may later use, in order to disclose any interesting patterns and quirks in the data.

The following sequence of table shells were designed before conducting a case-control study of Kawasaki syndrome. Kawasaki syndrome is a pediatric disease of unknown etiology which occasionally occurs in clusters. Two hypotheses to be tested by the case-control study were the syndrome’s association with antecedent viral illness and with recent exposure to carpet shampoo. A previously reported association with increasing household income was also to be evaluated.

Table Shell 1
Clinical features of Kawasaki syndrome cases with onset October–December, 1984

Clinical Feature	# with Feature	Percent
1. Fever ≥ 5 days	____	( )
2. Bilateral conjunctival injection	____	( )
3. Oral changes		
• injected lips	____	( )
• injected pharynx	____	( )
• dry, fissured lips	____	( )
• strawberry tongue	____	( )
4. Peripheral extremely changes		
• edema	____	( )
• erythema	____	( )
• periungual desquamation	____	( )
5. Rash	____	( )
6. Cervical lymphadenopathy <1.5 cm	____	( )
Total	____	**(100)**

Table Shell 2
Demographic characteristics of Kawasaki syndrome cases with onset October–December, 1984

Demographic characteristic		Number	Percent
Age	<1 yr	____	( )
	1 yr	____	( )
	2 yr		
	3 yr	____	( )
	4 yr	____	( )
	5 yr	____	( )
	≥6yr	____	( )
Sex	Male		
	Female	____	( )
Race	White	____	( )
	Black	____	( )
	Asian	____	( )
	Other	____	( )
Total		____	**(100)**

Alternatively, Table Shell 2 could have been drawn as a 3-variable table of number of cases by age by sex by race.

Figure 4.1
Illustration of table shells designed before conducting a case-control study of Kawasaki syndrome

Table Shell 3

County of Residence	Number	%
	___	( )
	___	( )

Table Shell 4

Household Income ($)	Number	%
≤ 10,000	___	( )
10,001 - 15,000	___	( )
15,001 - 20,000	___	( )
20,001 - 25,000	___	( )
25,001 - 20,000	___	( )
30,001 - 35,000	___	( )
≥ 35,000	___	( )

Table Shell 5

Number of Days in Hospital	Number	%
0	___	( )
1	___	( )
2	___	( )
3	___	( )
4	___	( )
5	___	( )

	Mean = ___
	Median = ___	

Table Shell 6

Serious Complication	Number	%
cardiac	___	( )
arthritis	___	( )
death	___	( )
	___	( )

Table Shell 7

Demographic Characteristic		Cases		Controls	
		Number	%	Number	%
AGE	< 1 yr	___	( )	___	( )
	1 yr	___	( )	___	( )
	2 yr	___	( )	___	( )
	3 yr	___	( )	___	( )
	4 yr	___	( )	___	( )
	5 yr	___	( )	___	( )
	6 yr	___	( )	___	( )
SEX	Male	___	( )	___	( )
	Female	___	( )	___	( )
RACE	White	___	( )	___	( )
	Black	___	( )	___	( )
	Asian	___	( )	___	( )
	Other	___	( )	___	( )
Total		___	( )	___	( )

Table Shell 8

Household Income ($)	Cases		Controls	
	Number	%	Number	%
≤ 10,000	___	( )	___	( )
10,001 - 15,000	___	( )	___	( )
15,001 - 20,000	___	( )	___	( )
20,001 - 25,000	___	( )	___	( )
25,001 - 20,000	___	( )	___	( )
30,001 - 35,000	___	( )	___	( )
≥ 35,000	___	( )	___	( )

Table Shell 3: Distribution by county of residence; Table Shell 4: Distribution by household income; Table Shell 5: Number of days of hospitalization; Table Shell 6: Distribution by serious complications; Table Shell 7: Demographic characteristics; and Table Shell 8: Household income.

Table Shell 9
Epidemiologic characteristics of Kawasaki syndrome cases and controls, with onset October–December, 1984

		Cases		Controls	
Epidemiologic characteristic		**Number**	**Percent**	**Number**	**Percent**
Antecedent illness	Yes	____	( )		( )
	No	____	( )		( )
		Odds ratio = ____, 95% CI = (,) x^2 = ____, p-value = ____			
Carpet shampoo exposure	Yes	____	( )		( )
	No	____	( )		( )
		Odds ratio = ____, 95% CI = (,) x^2 = ____, p-value = ____			

The sequence of table shells shown above and in Figure 4.1 provides a systematic, logical approach to the analysis. Of course, once the data are available and plugged into these tables, additional analyses will come to mind and should be pursued.

Creating Class Intervals

Some variables such as sex or “ate potato salad?” have a limited number of possible responses. These responses provide convenient categories for use in a table. When you study variables with a broader range of possible responses, such as time or systolic blood pressure, you must group the responses into a manageable number of categories (class intervals). In creating class intervals, keep the following guidelines in mind:

- Create class intervals that are mutually exclusive and that include all of the data. For example, if your first interval is 0-5, begin the next interval with 6, not 5. Also, consider what the **true** limits are. The true upper limit of 0.5 is 5.4999... for most measures, but 5.999... for age. True limits were discussed in Lesson 3.
- Use a relatively large number of narrow class intervals for your initial analysis. You can always combine intervals later. In general, you will wind up with 4 to 8 intervals.
- Use natural or biologically meaningful intervals when possible. Try to use age groupings that are standard or are used most frequently in the particular field of study. If rates are to be calculated, the intervals for the numerator must be the same as the intervals used for the available population data.
- Create a category for unknowns. For example, in the standard age groupings shown in Table 4.8 the categories created for unknowns are “age not stated,” “unknown,” and “not stated.”

Table 4.8 shows age groups commonly used by CDC for different purposes.

Table 4.8
Some standard groupings used at CDC

Notifiable diseases	**Pneumonia & influenza mortality**	**Final mortality statistics**	**HIV/AIDS**
<1 year	<28 days	<1 year	<5 years
1-4	28 days-<1 year	1-4	5-12
5-9	1-14	5-14	13-19
10-14	15-24	15-24	20-24
15-19	25-44	25-34	25-29
20-24	45-64	35-44	30-34
25-29	65-74	45-54	35-39
30-39	75-84	55-64	40-44
40-49	≥85	65-74	45-49
50-59	Unknown	75-84	50-54
≥60		≥85	55-59
Age not stated		Not stated	60-64
			≥65
Total	**Total**	**Total**	**Total**

Source: 3, 4, 21

Keep a natural baseline group as a separate category, even if the rest of the distribution has no natural distinctions. For example, in creating categories for cigarette smoking in cigarettes per day, leave nonsmokers (0 cigarettes/day) as a separate category and group smokers according to any of the arbitrary methods described below.

If no natural or standard class intervals are apparent, several strategies are available for creating intervals. Three strategies are described below.

Strategy 1: Divide the data into groups of similar size

Using this strategy, you set out to create a manageable number of class intervals, with about the same number of observations in each interval. Initially, you might use 8 intervals, collapsing them later into 4 for presenting the data to others. In effect, the 4 intervals represent the 4 quartiles of the data distribution. This method is well-suited to creating categories for area maps.

To apply this strategy, divide your total number of observations by the number of intervals you wish to create. Next, develop a cumulative frequency column of a rank-ordered distribution of your data to find where each interval break would fall.

Strategy 2: Base intervals on mean and standard deviation

With this strategy, you can create 3, 4, or 6 class intervals. To use this strategy, you must first find the mean and standard deviation of your distribution. (Lesson 3 covers the calculation of these measures.) You then use the mean plus or minus different multiples of the standard deviation to establish the upper limits for your intervals:

Upper limit of interval 1 = mean –2 standard deviations
Upper limit of interval 2 = mean –1 standard deviation
Upper limit of interval 3 = mean
Upper limit of interval 4 = mean +1 standard deviation
Upper limit of interval 5 = mean +2 standard deviations
Upper limit of interval 6 = maximum value

For example, suppose you wanted to establish six intervals for data that had a mean of 50 and a standard deviation of 10. The minimum value was 19; the maximum value was 82. You would calculate the upper limits of the six intervals as follows:

Upper limit of interval 1 = 50 – 20 = 30
Upper limit of interval 2 = 50 – 10 = 40
Upper limit of interval 3 = 50
Upper limit of interval 4 = 50 + 10 = 60
Upper limit of interval 5 = 50 + 20 = 70
Upper limit of interval 6 = maximum value = 82

If you then select the obvious lower limit for each upper limit, you have your six intervals:

Interval 1 = 19 – 30
Interval 2 = 31 – 40
Interval 3 = 41 – 50
Interval 4 = 51 – 60
Interval 5 = 61 – 70
Interval 6 = 71 – 82

You can create three or four intervals by combining some of the adjacent six-interval limits:

Six Intervals	Four Intervals	Three Intervals
Interval 1 = 19 – 30		
	Interval 1 = 19 – 40	Interval 1 = 19 – 40
Interval 2 = 31 – 40		
Interval 3 = 41 – 50	Interval 2 = 41 – 50	
		Interval 2 = 41 – 60
Interval 4 = 51 – 60	Interval 3 = 51 – 60	
Interval 5 = 61 – 70		
	Interval 4 = 61 – 82	Interval 3 = 61 – 82
Interval 6 = 71 – 82		

Strategy 3: Divide the range into equal class intervals

This method is the simplest and most commonly used, and is most readily adapted to graphs. To apply this method, do the following:

1. Find the range of the values in your data set. That is, find the difference between the maximum value (or some slightly larger convenient value) and zero (or the minimum value).
2. Decide how many class intervals (groups or categories) you want to have. For tables, we generally use 4 to 8 class intervals. For graphs and maps, we generally use 3 to 6 class intervals. The number will depend on what aspects of the data you want to disclose.
3. Find what size of class interval to use by dividing the range by the number of class intervals you have decided on.
4. Begin with the minimum value as the lower limit of your first interval and specify class intervals of whatever size you calculated until you reach the maximum value in your data.

Table 4.9
Mean annual age-adjusted cervical cancer mortality rates per 100,000 population, in rank order by state, United States, 1984-1986

Rank	State	Rate per 100,000	Rank	State	Rate per 100,000
1	SC	5.6	26	KS	3.6
2	WV	5.6	27	AR	3.6
3	AL	5.4	28	MD	3.5
4	LA	5.4	29	IA	3.4
5	AK	5.1	30	PA	3.4
6	TN	4.9	31	FL	3.4
7	ND	4.9	32	HI	3.4
8	KY	4.8	33	OR	3.3
9	MS	4.7	34	MI	3.3
10	NC	4.6	35	CA	3.2
11	GA	4.6	36	ID	3.1
12	ME	4.6	37	AZ	3.1
13	VT	4.3	38	MA	2.9
14	DE	4.3	39	NM	2.9
15	NH	4.3	40	WA	2.8
16	IN	4.1	41	NV	2.8
17	OK	4.1	42	CT	2.8
18	IL	4.0	43	RI	2.8
19	MT	4.0	44	WI	2.7
20	VA	3.9	45	CO	2.5
21	OH	3.8	46	NE	2.4
22	MO	3.8	47	SD	2.4
23	TX	3.7	48	MN	2.2
24	NY	3.7	49	WY	1.9
25	NJ	3.7	50	UT	1.8
			Total	**U.S.**	**3.7**

Source: 2

Example

In the example, we will demonstrate each strategy for creating categories using the cervical cancer mortality rates shown in Table 4.9. In each case, we will create four class intervals of rates.

Strategy 1: Divide the data into groups of similar size

(Note: If Table 4.9 had been arranged alphabetically, the first step would have been to sort the data into rank order by rate. Fortunately, this has already been done.)

1. Divide the list into four equal-sized groups of places:
   50 states ÷ 4 = 12.5 states per group. Because we can't cut a state in half, we will have to use two groups of 12 states and two groups of 13 states. Since Vermont (#13) could go into either the first or second group and Massachusetts (#38) could go into either third or fourth group, we create the following groups:

   a. South Carolina through Maine (1 through 12)
   b. Vermont through New Jersey (13 through 25)
   c. Kansas through Arizona (26 through 37)
   d. Massachusetts through Utah (38 through 50)

   Notice that this arrangement puts Vermont with Delaware (both have rates of 4.3), and puts Massachusetts with New Mexico (both have rates of 1.8).

2. Identify the rate for the first and last state in each group:

States	Rates per 100,000
a. ME–SC	4.6–5.6
b. NJ–VT	3.7–4.3
c. AZ–KS	3.1–3.6
d. UT–MA	1.8–2.9

3. Adjust the limits of each interval so no gap exists between the end of one class interval and beginning of the next (compare the intervals below with those above):

States	Rates per 100,000	Number of states
a. ME–SC	**4.5**–5.6	12
b. NJ–VT	3.7–**4.4**	13
c. AZ–KS	**3.0**–3.6	12
d. UT–MA	1.8–**2.9**	13

Strategy 2: Base intervals on mean and standard deviation

1. Calculate the mean and standard deviation (Lesson 3 describes how to calculate these measures.):

   Mean = 3.70
   Standard deviation = 0.96

2. Find the upper limits of 4 intervals (Note: We demonstrated creating 4 intervals by first creating 6 intervals and then combining the upper and lower pairs of intervals. Here, however, we will simply use the appropriate upper limit of the pairs that would be combined.)

   Upper limit of interval 1: mean – 1 standard deviation = 2.74
   Upper limit of interval 2: mean = 3.70
   Upper limit of interval 3: mean + 1 standard deviation = 4.66
   Upper limit of interval 4: maximum value = 5.6

3. Select the lower limit for each upper limit to define four full intervals. Specify the states that fall into each interval (Note: To place the states with the highest rates first we have reversed the order of the intervals):

States	Rates per 100,000	Number of states
a. MS–SC	4.67–5.60	9
b. MO–NC	3.71–4.66	13
c. RI–TX	2.75–3.70	21
d. UT–WI	1.80–2.74	7

Strategy 3: Divide the range into equal class intervals

1. Divide the range from zero (or the minimum value) to the maximum by 4:

   (5.6 – 1.8) / 4 = 3.8 / 4 = 0.95

2. Use multiples of 0.95 to create four categories, starting with 1.8:

   1.80 through (1.8 + 0.95) = 1.8 through 2.75
   2.76 through (1.8 + 2 × 0.95) = 2.76 through 3.70
   3.71 through (1.8 + 3 × 0.95) = 3.71 through 4.65
   4.66 through (1.8 + 4 × 0.95) = 4.66 through 5.6

3. Final categories:

States	Rates per 100,000	Number of states
a. MS–SC	4.66–5.60	9
b. MO–NC	3.71–4.65	13
c. RI–TX	2.76–3.70	21
d. UT–WI	1.80–2.75	7

4. Alternatively, since 0.95 is close to 1.0, multiples of 1.0 might be used to create the four categories. Start at the center value (5.6 + 1.8)/2 = 3.7, subtract 1.0 to determine the upper limit of the first interval (2.7). The upper limits of the third and fourth intervals will be 3.7 + 1.0 = 4.7, and 3.7 + 2 × 1.0 = 5.7.

Final categories:

States	Rates per 100,000	Number of states
a. KY–SC	4.71–5.70	8
b. MO–MS	3.71–4.70	14
c. RI–TX	2.71–3.70	21
d. UT–WI	1.71–2.70	7

Exercise 4.2

With the data on cervical cancer mortality rates presented in Table 4.9, use each strategy to create **three** class intervals for the rates.

Answers on page 270.

Table 4.9, revisited
Mean annual age-adjusted cervical cancer mortality rates per 100,000 population, in rank order by state, United States, 1984-1986

Rank	State	Rate per 100,000	Rank	State	Rate per 100,000
1	SC	5.6	26	KS	3.6
2	WV	5.6	27	AR	3.6
3	AL	5.4	28	MD	3.5
4	LA	5.4	29	IA	3.4
5	AK	5.1	30	PA	3.4
6	TN	4.9	31	FL	3.4
7	ND	4.9	32	HI	3.4
8	KY	4.8	33	OR	3.3
9	MS	4.7	34	MI	3.3
10	NC	4.6	35	CA	3.2
11	GA	4.6	36	ID	3.1
12	ME	4.6	37	AZ	3.1
13	VT	4.3	38	MA	2.9
14	DE	4.3	39	NM	2.9
15	NH	4.3	40	WA	2.8
16	IN	4.1	41	NV	2.8
17	OK	4.1	42	CT	2.8
18	IL	4.0	43	RI	2.8
19	MT	4.0	44	WI	2.7
20	VA	3.9	45	CO	2.5
21	OH	3.8	46	NE	2.4
22	MO	3.8	47	SD	2.4
23	TX	3.7	48	MN	2.2
24	NY	3.7	49	WY	1.9
25	NJ	3.7	50	UT	1.8
			Total	**U.S.**	**3.7**

Source: 2

Graphs

A graph is a way to show quantitative data visually, using a system of coordinates. It is a kind of statistical snapshot that helps us see patterns, trends, aberrations, similarities, and differences in the data. Also, a graph is an ideal way of presenting data to others. Your audience will remember the important aspects of your data better from a graph than from a table.

In epidemiology, we commonly use rectangular coordinate graphs, which have two lines, one horizontal and one vertical, that intersect at a right angle. We refer to these lines as the horizontal axis (or *x-axis*), and the vertical axis (or *y-axis*). We usually use the horizontal axis to show the values of the **independent (or *x*) variable**, which is the method of classification, such as time. We use the vertical axis to show the **dependent (or *y*) variable**, which, in epidemiology, is usually a frequency measure, such as number of cases or rate of disease. We label each axis to show what it represents (both the name of the variable and the units in which it is measured) and mark a scale of measurement along the line.

Table 4.10 shows the number of measles cases by year of report from 1950 to 1989. We have used a portion of these data to create the graph shown in Figure 4.2. The independent variable, years, is shown on the horizontal axis. The dependent variable, number of cases, is shown on the vertical axis. A grid is included in Figure 4.2 to illustrate how points are plotted. For example, to plot the point on the graph for the number of cases in 1953, draw a line up from 1953, then draw a line from 449 cases to the right. The point where these lines intersect is the point for 1953 on the graph. By using the data in Table 4.10, complete the graph in Figure 4.2 by plotting the points for 1955 to 1959.

Arithmetic-scale Line Graphs

An arithmetic-scale line graph shows patterns or trends over some variable, usually time. In epidemiology, we commonly use this type of graph to show a long series of data and to compare several series. It is the method of choice for plotting rates over time.

In an arithmetic-scale line graph, a set distance along an axis represents the same quantity anywhere on that axis. This holds true for both the *x-axis* and the *y-axis*. In Figure 4.3, for example, the space between tick marks along the *y-axis* represents an increase of 100,000 (100 x 1000) cases anywhere along the axis.

Several series of data can be shown on the same arithmetic-scale line graph. In Figure 4.4, one line represents the decline of rabies in domestic animals since 1955, while another line represents the concurrent rise of rabies in wild animals. A third line represents the total.

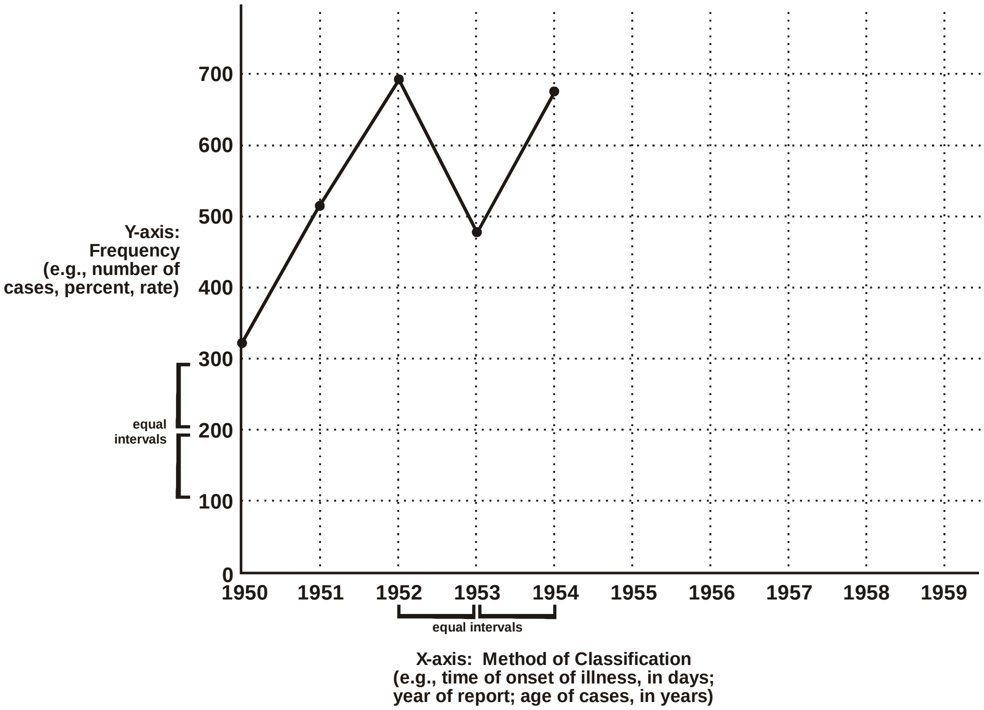
What scale we use on the *x-axis* depends on what intervals we have used for our independent variable in collecting the data. Usually, we plot time data with the same specificity we use to collect them, e.g., weekly, annually, and so forth. If we have used very small intervals in collecting the data, however, we can easily collapse those intervals into larger ones for displaying the data graphically.

Table 4.10
Measles (rubeola) by year of report, United States, 1950-1989

Year	Reported cases (x1,000)	Year	Reported cases (x1,000)
1950	319	1970	47
1951	530	1971	75
1952	683	1972	32
1953	449	1973	27
1954	683	1974	22
1955	555	1975	24
1956	612	1976	41
1957	487	1977	57
1958	763	1978	27
1959	406	1979	14
1960	442	1980	13
1961	424	1981	3
1962	482	1982	2
1963	385	1983	1
1964	458	1984	3
1965	262	1985	3
1966	204	1986	6
1967	63	1987	4
1968	22	1988	3
1969	26	1989	18

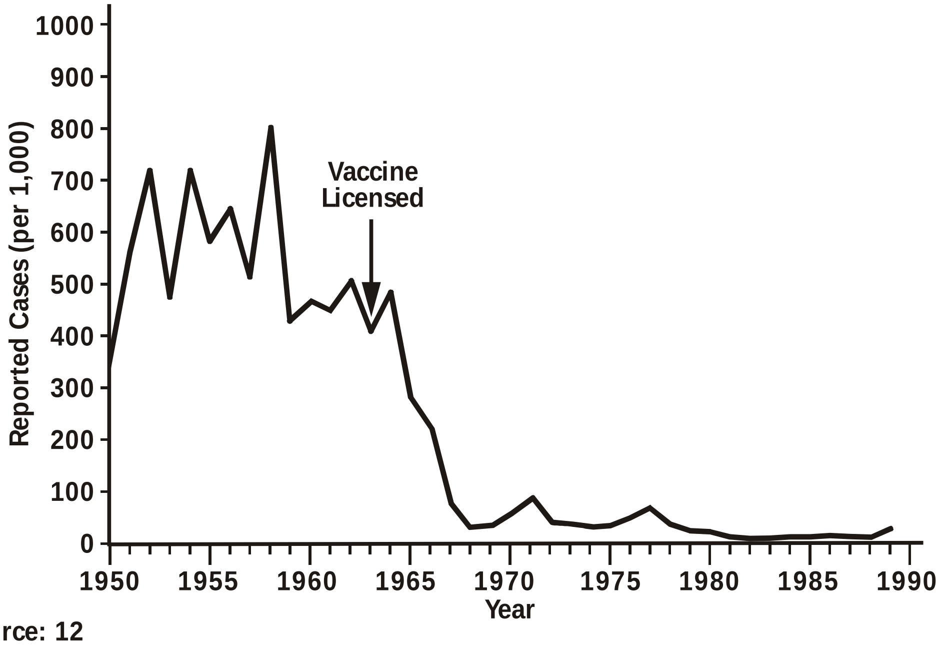
Source: 12

Figure 4.2
Partial graph of measles (rubeola) by year of report, United States, 1950-1959



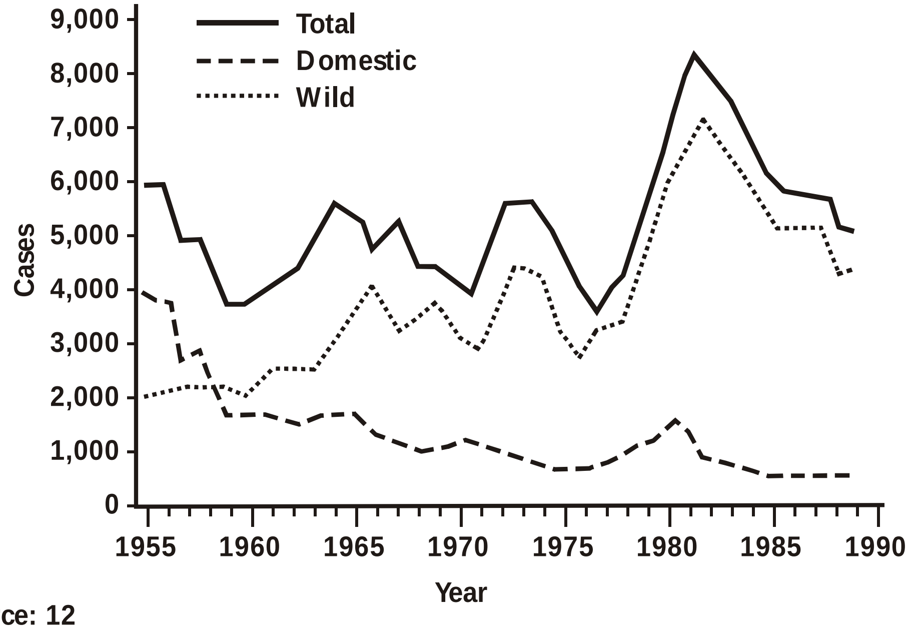
Source: 12

Figure 4.3
Example of arithmetic-scale line graph:
Measles (rubeola) by year of report, United States, 1950-1989



Source: 12

Figure 4.4
Example of arithmetic-scale line graph:
Rabies, wild and domestic animals by year of report,
United States and Puerto Rico, 1955-1989



Source: 12

To select a scale for the *y-axis*, do the following:

- Make the *y-axis* shorter than the *x-axis*, so that your graph is horizontal (i.e., the horizontal length is greater than the vertical length), and make the two axes in good proportion: an x:y ratio of about 5:3 is often recommended.
- Always start the *y-axis* with 0.
- Determine the range of values you need to show on the *y-axis* by identifying the largest value you need to graph on the *y-axis* and rounding that figure off to a number slightly larger than that. For example, the largest y-value in Figure 4.3 is 763,094 in 1958. This value was rounded up to 1,000,000 for determining the range of values that were shown on the *y-axis*.
- Select an interval size that will give you enough intervals to show the data in enough detail for your purposes. In Figure 4.3, 10 intervals of 100,000 each were considered adequate to show the important details of the data.
- If the range of values to show on the *y-axis* includes a gap, that is, an area of the graph that will have no data points, a scale break may be appropriate. With a scale break the *y-axis* stops at the point where the gap begins and starts again where the gap ends. Scale breaks should be used only with scale line graphs.

Exercise 4.3

In both graphs, be sure to use intervals on the y-axis that are appropriate for the range of data you are graphing. Graph paper is provided in Appendix D.

A. Construct an arithmetic-scale line graph of the measles data in Table 4.11, showing measles rates from 1955-1990 with a single line.

B. Construct an arithmetic-scale line graph of the measles data for 1980-1990.

Table 4.11
Measles (rubeola) rate per 100,000 population,
United States, 1955-1990

Year	Rate	Year	Rate	Year	Rate
1955	336.3	1967	31.7	1979	6.2
1956	364.1	1968	11.1	1980	6.0
1957	283.4	1969	12.8	1981	1.4
1958	438.2	1970	23.2	1982	0.7
1959	229.3	1971	36.5	1983	0.6
1960	246.3	1972	15.5	1984	1.1
1961	231.6	1973	12.7	1985	1.2
1962	259.0	1974	10.5	1986	2.6
1963	204.2	1975	11.4	1987	1.5
1964	239.4	1976	19.2	1988	1.4
1965	135.1	1977	26.5	1989	7.3
1966	104.2	1978	12.3	1990	10.7

Source: 12

Answer on page 272.

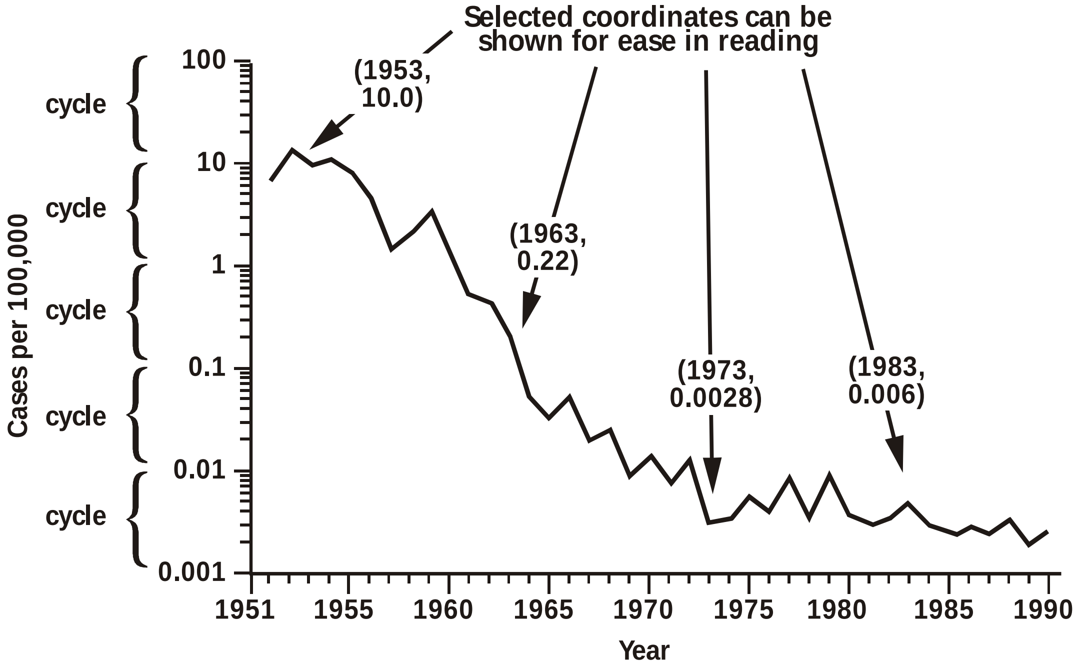
Semilogarithmic-scale Line Graphs

In a semilogarithmic-scale line graph (a “semilog graph”), the divisions on the *y-axis* are logarithmical, rather than arithmetical as on arithmetic-scale line graphs. The *x-axis* has an arithmetic scale, as it does on arithmetic-scale line graphs.

Figure 4.5 shows an example of a semilog graph. Notice the following characteristics of the scale on the y-axis:

- There are five cycles of tick-marks along the axis; each cycle covers equal distance on the scale.
- Each cycle represents one order of magnitude greater than the one below it, that is, the values in each cycle are ten times greater than those in the preceding one. Notice, for example, that the values in the 4th cycle are 1 to 10 and in the 5th are 10 to 100 but the distances on the scale are the same.
- Within a cycle are ten tick-marks, with the space between tick marks becoming smaller and smaller as they move up the cycle. Thus the distance from 1 to 2 is not the same as the distance from 2 to 3.
- The axis covers a large range of *y-values*, which might have been difficult to show clearly on an arithmetic scale. Semilog graphs are useful when you must fit a wide range of values on a single graph, as here.

Figure 4.5
Example of semilogarithmic-scale line graph:
Reported cases of paralytic poliomyelitis per 100,000 population
by year of occurrence, United States, 1951-1989



Source: 12

Because of the logarithmic scale, equal distances on the y-axis represent an equal percentage of change. This characteristic makes a semilog graph particularly useful for showing rates of change in data. To interpret data in a semilog graph, you must understand the following characteristics of the graph:

- A sloping straight line indicates a constant rate (not amount) of increase or decrease in the values.
- A horizontal line indicates no change.
- The slope of the line indicates the rate of increase or decrease.
- Two or more lines following parallel paths show identical rates of change.

Semilog graph paper is available commercially, and most include at least three cycles. To find how many cycles you need, do the following:

1. Find your smallest *y*-value and identify what order of magnitude it falls within. This establishes what your first cycle will represent.

   For example, if your smallest *y*-value is 47 your first cycle will begin with 10 and end with 100; if it is 352, your first cycle will begin with 100 and end with 1,000.

2. Find your largest *y*-value and identify what order of magnitude it falls within. This establishes what your last cycle will represent.

   For example, if your largest *y*-value is 134,826, your last cycle will begin with 100,000. Although a full cycle that begins with 100,000 ends with 1,000,000, you would not need to show the entire cycle. It would be sufficient to show only the first few tick-marks in your last cycle: 100,000, 200,000, and 300,000.

3. Identify how many cycles lie between your first and last cycles. You will need that number of cycles, plus two to include the first and last cycles.

   So, if your smallest y-value is 47, and your largest y-value is 134,826, you will need the following cycles:

   10-100
   100-1,000
   1,000-10,000
   10,000-100,000
   100,000-1,000,000

   Thus, with *y*-values ranging from 47 to 134,826, you will need four cycles and part of a fifth.

Figure 4.6
Possible values which could be assigned to the *y*-axis of a semilogarithmic-scale line graph

Possible Values			
20	2,000	20,000	200,000
10	1,000	10,000	100,000
8	800	8,000	80,000
6	600	6,000	60,000
5	500	5,000	50,000
4	400	4,000	40,000
3	300	3,000	30,000
2	200	2,000	20,000
1.0	100	1,000	10,000
.8	80	800	8,000
.6	60	600	6,000
.5	50	500	5,000
.4	40	400	4,000
.3	30	300	3,000
.2	20	200	2,000
.1	10	100	1,000
.08	8	80	800
.06	6	60	600
.05	5	50	500
.04	4	40	400
.03	3	30	300
.02	2	20	200
.01	1.0	10	100
.008	.8	8	80
.006	.6	6	60
.005	.5	5	50
.004	.4	4	40
.003	.3	3	30
.002	.2	2	20
0.001	0.1	1	10

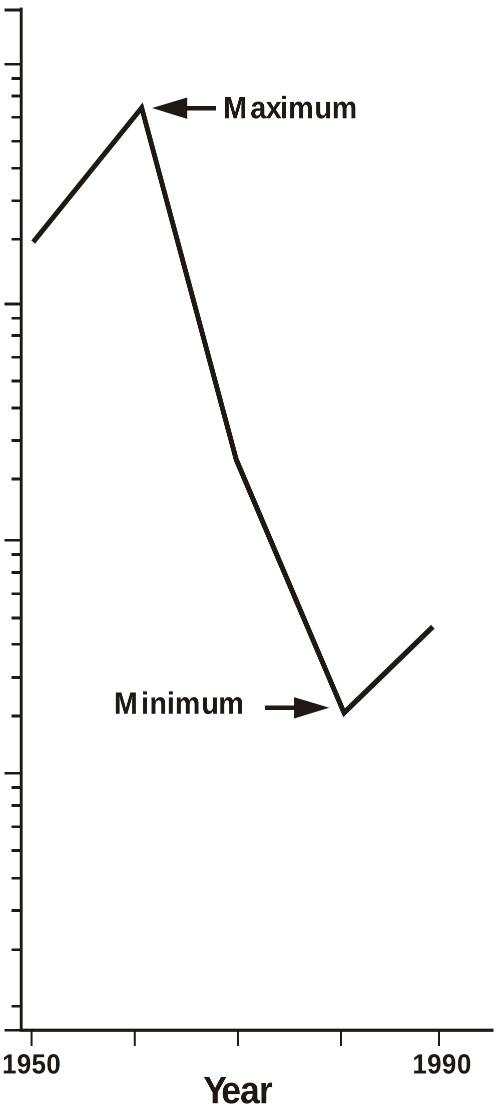


Figure 4.6 shows some of the ranges of values that could be shown on a four-cycle *y-axis* of a semilog graph.

The type of line graph you use depends primarily on whether you want to show the *actual changes* in a set of values or whether you want to emphasize *rates of change*. To show actual changes, use an arithmetic scale on the *y-axis* (an arithmetic-scale line graph). To show rates of change, use a logarithmic scale on the *y-axis* (a semilogarithmic-scale line graph). However, you might also choose a semilog graph—even when you are interested in actual changes in the data—when the range of the values you must show on the *y-axis* is awkwardly large.

Exercise 4.4

Graph the measles data in Table 4.11, page 231, with a semilogarithmic-scale line graph. Semilog graph paper with five cycles is provided in Appendix D.

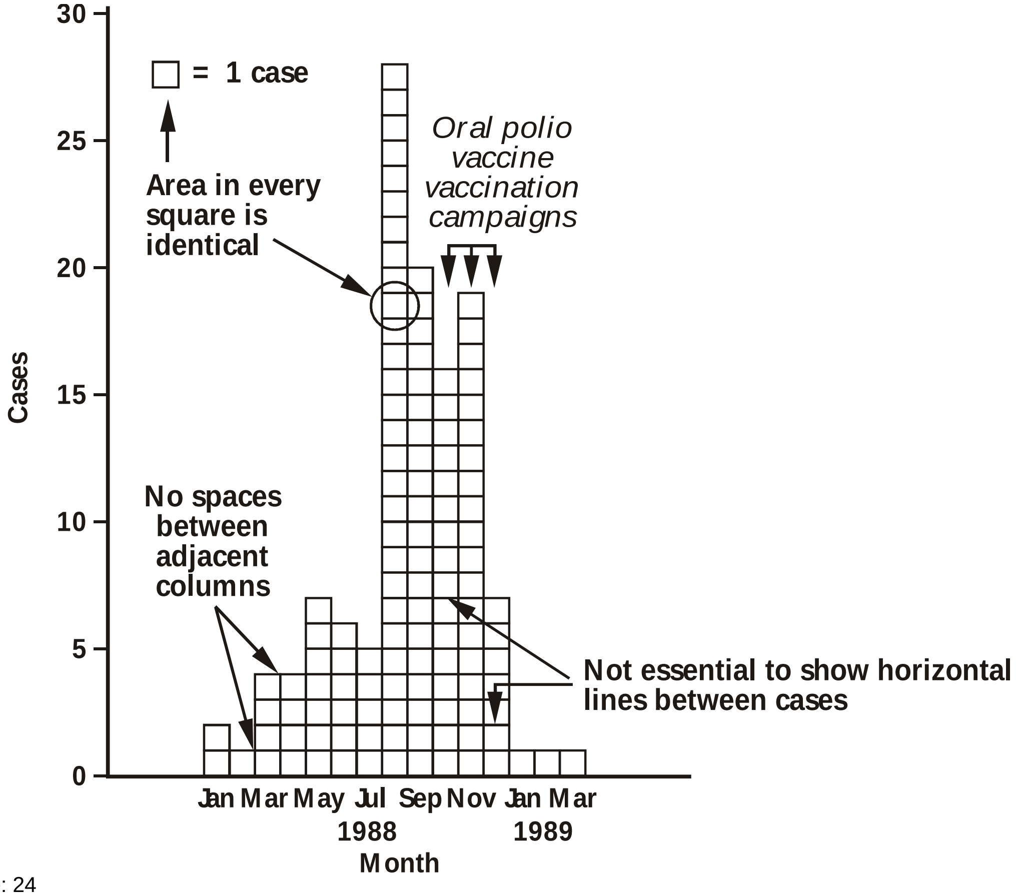
Answer on page 273.

Histograms

A histogram is a graph of the frequency distribution of a continuous variable. It uses adjoining columns to represent the number of observations for each class interval in the distribution. The *area* of each column is proportional to the number of observations in that interval.

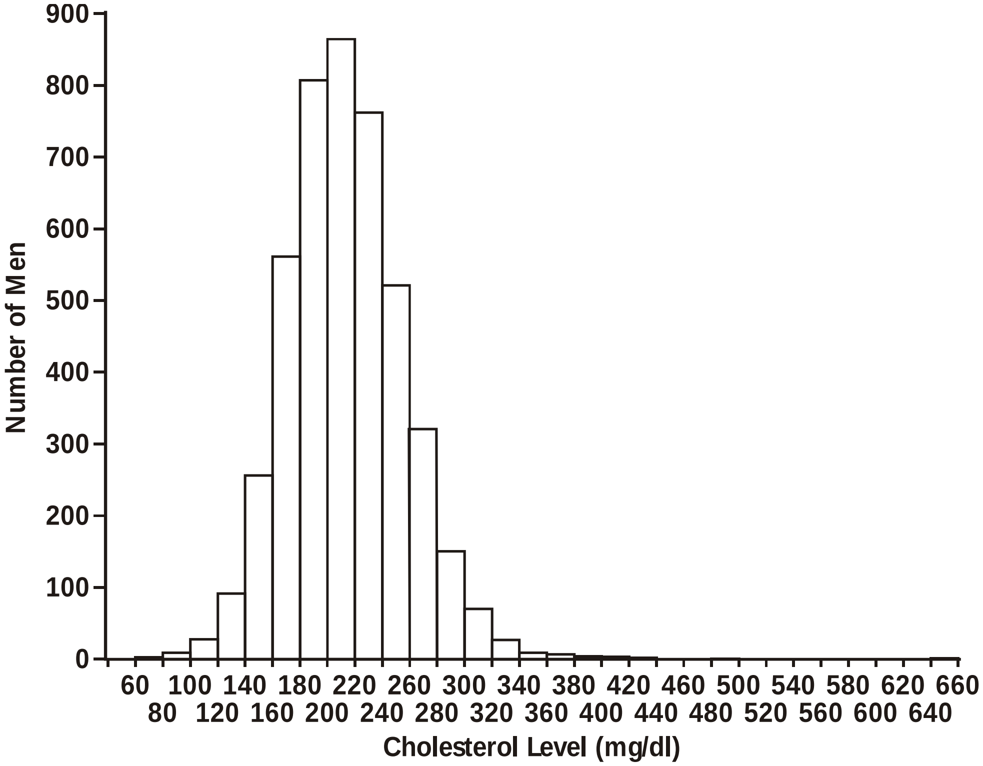
Figures 4.7, 4.8, and 4.9 show histograms of frequency distributions with equal class intervals. Since all class intervals are equal in these histograms, the height of each column is in proportion to the number of observations it depicts. Histograms with unequal class intervals are difficult to construct and interpret properly, and are not recommended. Neither should you use scale breaks in the *y-axis* of histograms, because they give a deceptive picture of relative frequencies.

Figure 4.7
Example of histogram: Reported cases of paralytic poliomyelitis by month of occurrence, Oman, January 1988-March 1989



Source: 24

Figure 4.8
Example of histogram: Reported cholesterol levels among 4,462 men, Men's Health Study, United States, 1985-1986



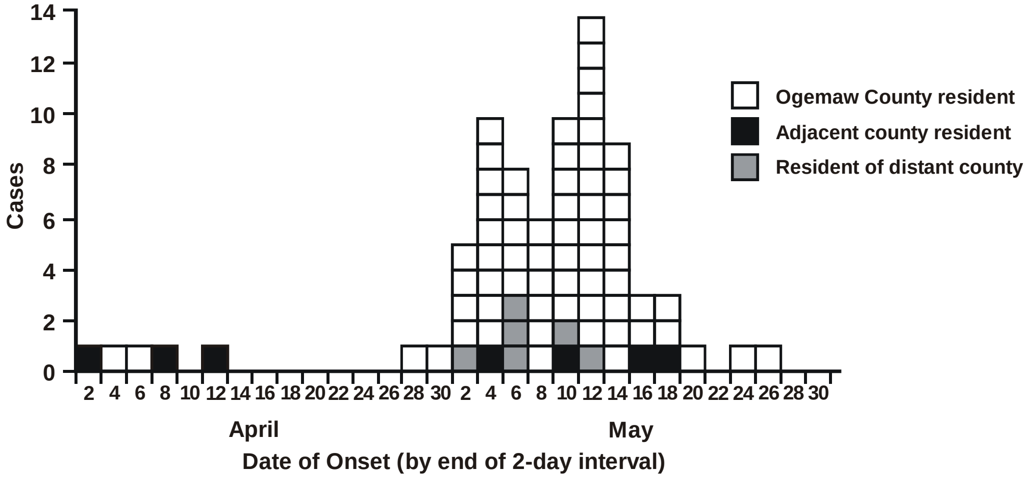
Source: 13

The most common *x-axis* variable is time, as shown in figures 4.7, 4.9, and 4.10. However, other continuous variables such as cholesterol level or blood pressure level may be used on the *x-axis*. Figure 4.8 shows the frequency of observations by cholesterol level in class intervals.

You may show a second variable with a histogram by shading each column into the component categories of the second variable. Suppose, for example, that we wanted to show the number of hepatitis A cases by date of onset and residency status. In Figure 4.9 the appropriate number of non-residents are shaded at the bottom of each column. When you show data in this format, however, it is difficult to compare the upper component from column to column because it does not have a flat baseline. Therefore, you should put the component that is of most interest at the bottom of the columns. Alternatively, instead of shading columns, you can create a separate histogram for each component of the second variable, stacking them for display, as in Figure 4.10.

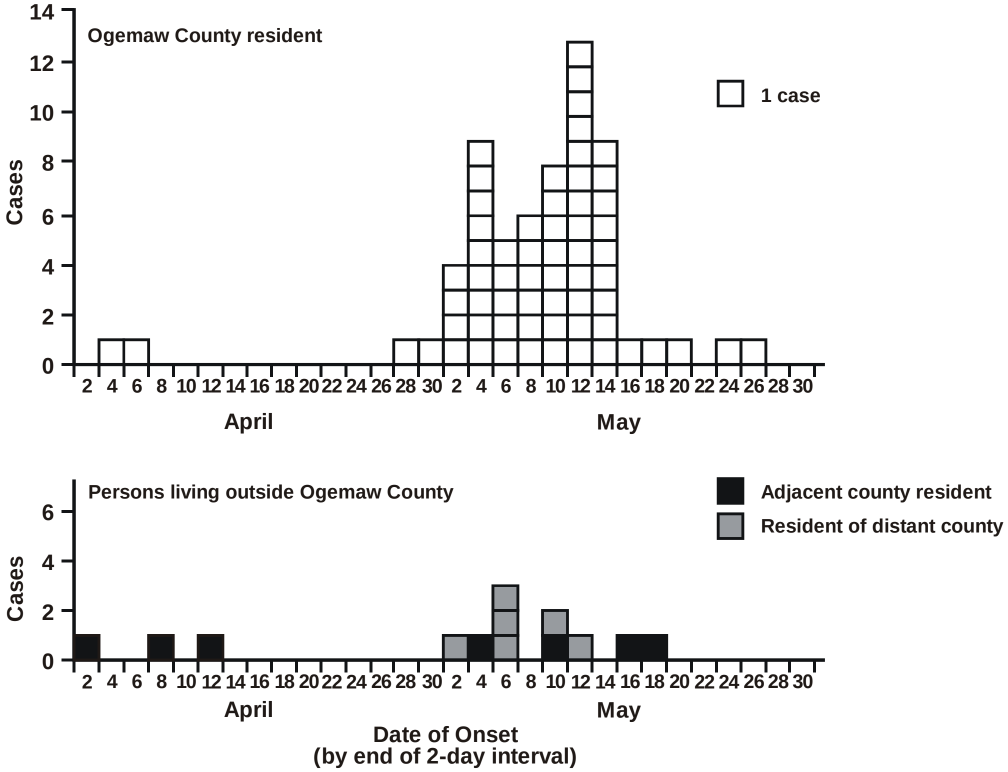
Compare Figures 4.9 and 4.10. They contain the same data, but in different formats. Which format do you prefer for comparing the time pattern of cases among residents and non-residents?

Figure 4.9
Example of histogram:
Number of reported cases of hepatitis A
by date of onset and residency status, Ogemaw County, April-May 1968



Source: 22

Figure 4.10
Example of histogram:
Number of reported cases of hepatitis A
by date of onset and residency status, Ogemaw County, April-May 1968



Source: 22

It is sometimes helpful to include a box or rectangle to show how many values of *y* (usually cases) that a given height of a column represents. We make this legend as wide as a single column, and as high as some convenient number of values on the *y-axis*—1, 5, 10, . . . etc. We note beside the square or rectangle what it represents, e.g., 1 case or 5 cases.

Epidemiologists frequently create and discuss *epidemic curves*. An epidemic curve isn't a curve at all, but a histogram that shows cases of disease during a disease outbreak or epidemic by their date of onset. As shown in Figure 4.9, we often draw the columns as stacks of squares, with each square representing one case. Figure 4.9 shows us that one person had the onset of symptoms between April 27 and 28, one more person had the onset on April 29 or 30, and between May 1 and 2 five additional individuals had the onset of symptoms. We show the duration of the epidemic along the *x-axis* in equal time periods. On an epidemic curve, each number should be centered between the tick marks of the appropriate interval. We use whatever interval of time is appropriate for the disease in question: perhaps hours for an outbreak of *C. perfringens* gastroenteritis, or 3-5 days for an outbreak of hepatitis A. As a general rule, we make the intervals less than one-fourth of the incubation period of the disease shown. We begin the *x-axis* before the first case of the outbreak, and show any cases of the same disease which occurred during the pre-epidemic period. These cases may represent background or unrelated cases. They may also represent the source of the outbreak!

Exercise 4.5

Using the data from the nursing home outbreak in Exercise 4.1 (see page 213), draw an epidemic curve. Describe the features of this graph as if you were speaking over the telephone to someone who cannot see the graph. Graph paper is provided in Appendix D.

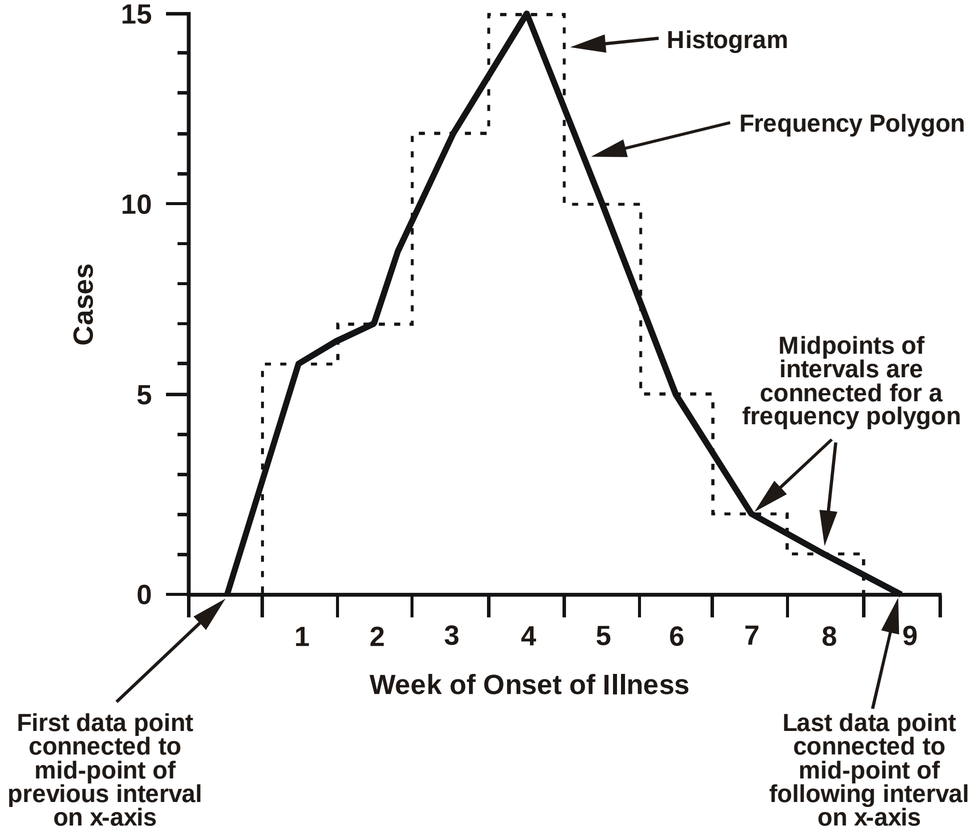
Answer on page 274.

Frequency Polygons

A frequency polygon, like a histogram, is the graph of a frequency distribution. In a frequency polygon, we mark the number of observations within an interval with a single point placed at the midpoint of the interval, and then connect each set of points with a straight line. Figure 4.11 shows an example of a frequency polygon over the outline of a histogram for the same data. Ordinarily, we wouldn't show both on the same graph. Showing both here, however, lets you compare their construction.

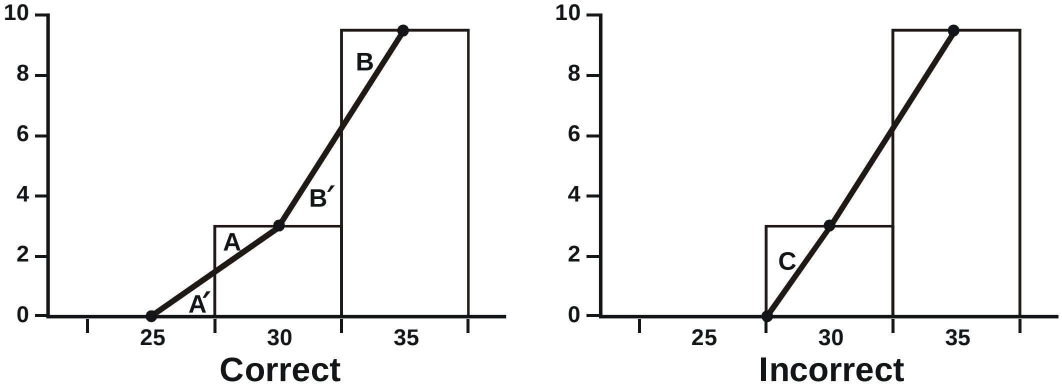
Notice how the histogram and the line of the frequency polygon—as it moves from midpoint to midpoint—create a series of equal-sized pairs of triangles—one that lies outside the histogram and one that lies inside it. This is a necessary aspect of frequency polygons: a frequency polygon of a set of data must enclose the same area as a histogram of that data: for every area of histogram that the polygon leaves out, it must import an area of equal size.

Figure 4.11
Number of reported cases of influenza-like illness by week of onset



To maintain an equal total area you must pay special attention to how you "close" a frequency polygon. Figure 4.12 shows the correct method at the left and the incorrect method at the right—again superimposed on a corresponding segment of a histogram. In the correct method, notice that the line of the frequency polygon begins in the interval below the first interval that contains any observations, completely outside the histogram. It begins at the midpoint of that interval (with a *y* value of 0) and connects with the midpoint of the first interval that contains observations. This extension of the line beyond the values observed in the data serves to create an area A′ under the polygon that equals area A that is cut out of the corresponding histogram. Notice in Figure 4.11 that the right side of a frequency polygon is closed in a similar way.

Figure 4.12
Correct method of closing a frequency polygon at left; incorrect method for closing a frequency polygon at right

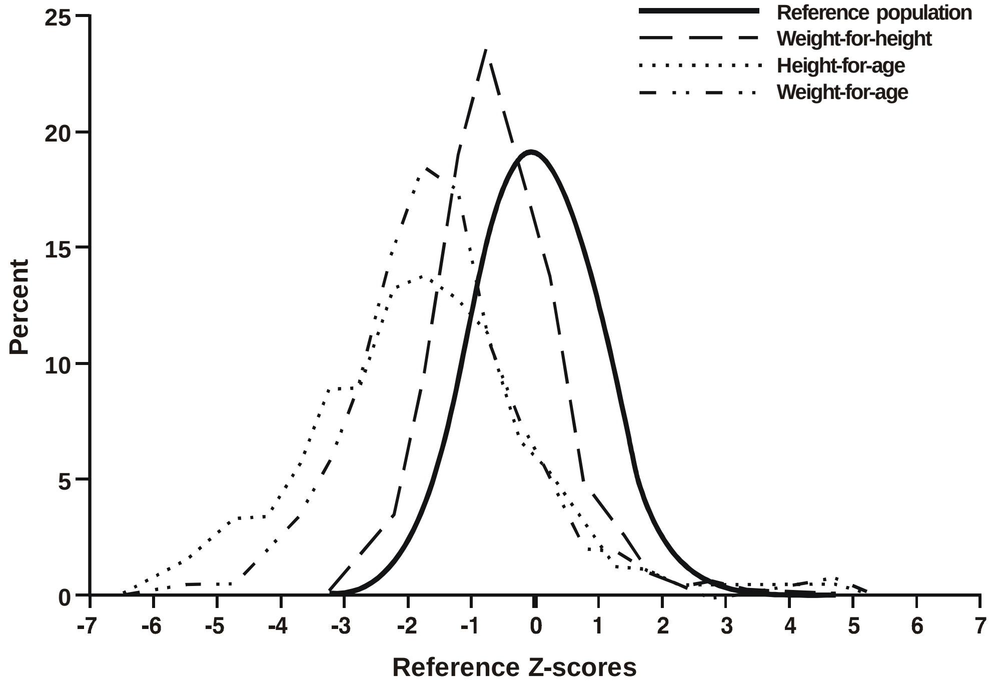


In contrast, the incorrect but unfortunately common method of closing a frequency polygon is shown at the right in Figure 4.12. Here, the line has been brought to the baseline at the beginning of the first interval that contains observations, cutting off an area, C, from inside the histogram without enclosing an equal area from outside the histogram. As a consequence, the area under the polygon would not be in proportion to the total number of observations in the data set.

Frequency polygons make it easy to depict and compare two or more distributions on the same set of axes. Figure 4.13 shows a graph in which three frequency polygons are compared with each other and to the normal distribution.

A frequency polygon differs from an arithmetic-scale line graph in several ways. We use a frequency polygon (or histogram) to display the entire frequency distribution (counts) of a continuous variable. We use an arithmetic-scale line graph to plot a series of observed data points (counts or rates), usually over time. A frequency polygon must be closed at both ends because the area under the curve is representative of the data; an arithmetic-scale line graph simply plots the data points.

Figure 4.13
Anthropometry of Haitian children ages 24.0 to 59.9 months compared with CDC's National Center for Health Statistics/World Health Organization reference population, northern departments of Haiti, 1990



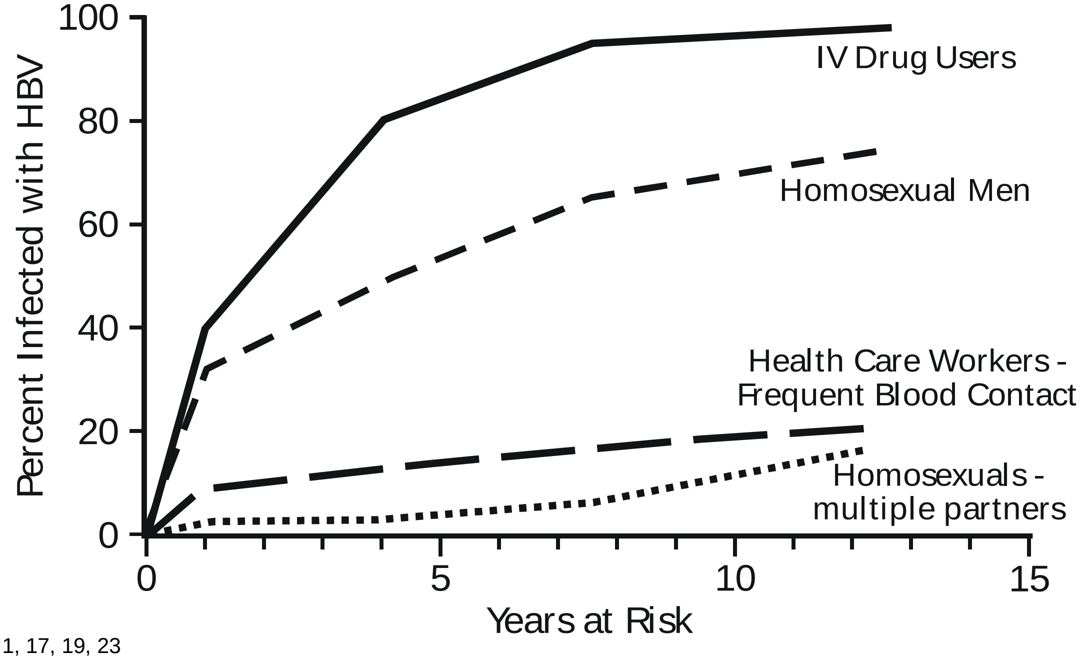
Source: 9

Cumulative Frequency and Survival Curves

As its name implies, a cumulative frequency curve plots the cumulative frequency rather than the actual frequency for each class interval of a variable. Figure 4.14 shows a graph with four cumulative frequency curves. This type of graph is useful for identifying medians, quartiles, and other percentiles. The *x-axis* records the class intervals and the *y-axis* shows the cumulative frequency either on an absolute scale (e.g., number of cases) or as proportions of 100%. We plot each cumulative frequency at the upper limit of the interval it applies to, rather than at the midpoint. This practice allows the graph to represent visually the number or percentage of observations above and below the particular value.

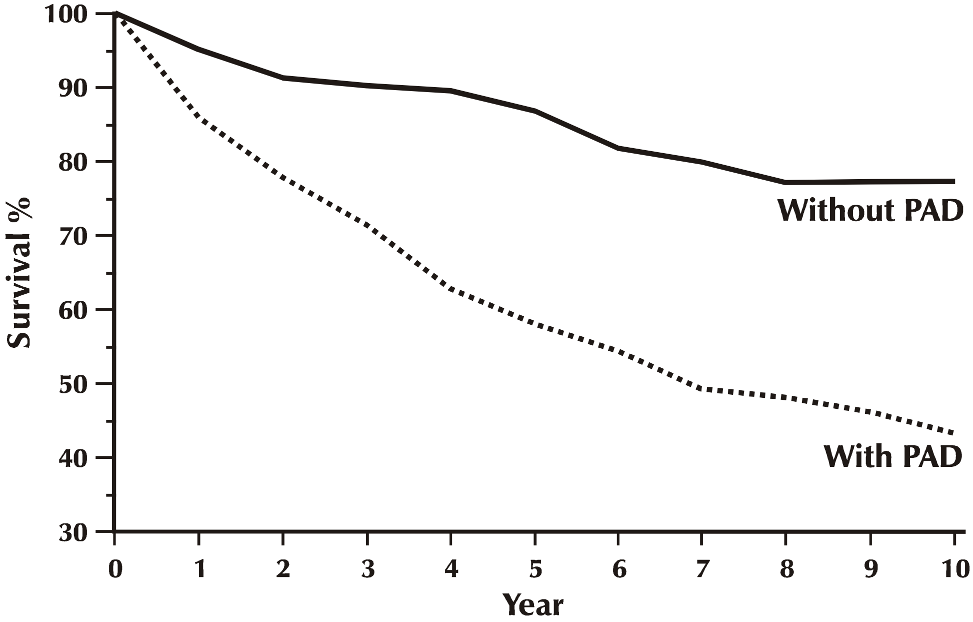
A survival curve is used with follow-up studies to display the proportion of one or more groups still alive at different time periods. Similar to the axes of the cumulative frequency curve, the *x-axis* records the time periods and the *y-axis* shows percentages, from 0% to 100%, still alive. The most striking difference is in the plotted curves themselves. Whereas a cumulative frequency starts at zero in the lower left corner of the graph and approaches 100% in the upper right corner, a survival curve begins at 100% in the upper left corner and proceeds toward the lower right corner as members of the group die. The survival curve in Figure 4.15 compares the percentage of survival by those with peripheral arterial disease (PAD) with those without PAD. Which group has the higher survival percentage (or survival experience)? By Year 10 the survival experience for those without PAD was substantially better than those with PAD.

Figure 4.14
Cumulative incidence of hepatitis B virus infection by duration of high-risk behavior



Source: 1, 17, 19, 23

Figure 4.15
Survival curves for a cohort of patients with peripheral arterial disease (PAD) (n=482) and without PAD (n=262) Pittsburgh, Pennsylvania, 1977-1985



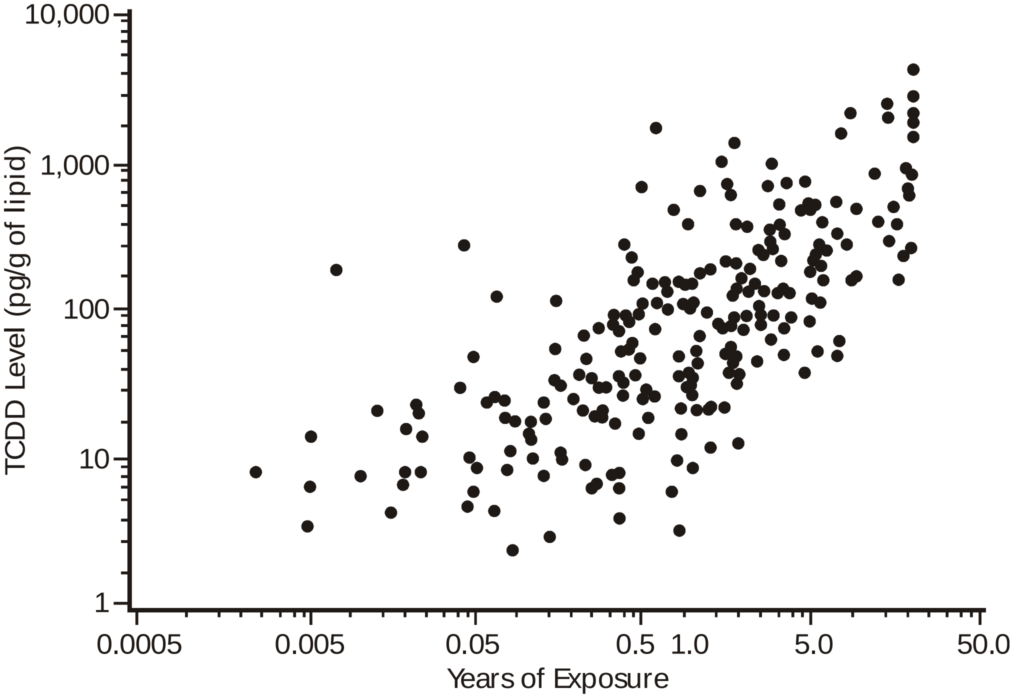
Source: 20

Scatter Diagrams

A scatter diagram (or “scattergram”) is a graph used for plotting the relationship between two continuous variables, with the *x-axis* representing one variable and the *y-axis* representing the other. To create a scatter diagram we must have a pair of values for every person, group, or other entity in our data set, one value for each variable. We then plot each pair of values by placing a point on the graph where the two values intersect. Figure 4.16 shows a scatter diagram that plots serum tetrachlorodibenzo-*p*-dioxin (TCDD) levels by years of exposure for a group of workers.

In interpreting a scatter diagram, we look at the overall pattern made by the plotted points. A fairly compact pattern indicates a high degree of correlation. Widely scattered points indicate little correlation. If we want a more exact, quantitative measure of the relationship between the variables in a scatter diagram, we can use formal statistical methods, such as linear regression. We will not cover those methods in this course.

Figure 4.16
Example of scattergram:
Serum levels of tetrachlorodibenzo-*p*-dioxin (TCDD),
as adjusted for lipids, in 253 workers, according to years
of exposure, 12 chemical plants, United States, 1987



Source: 16

Charts

Charts are methods of illustrating statistical information using only **one** coordinate. They are most appropriate for comparing data with discrete categories other than place, but have many other uses as well.

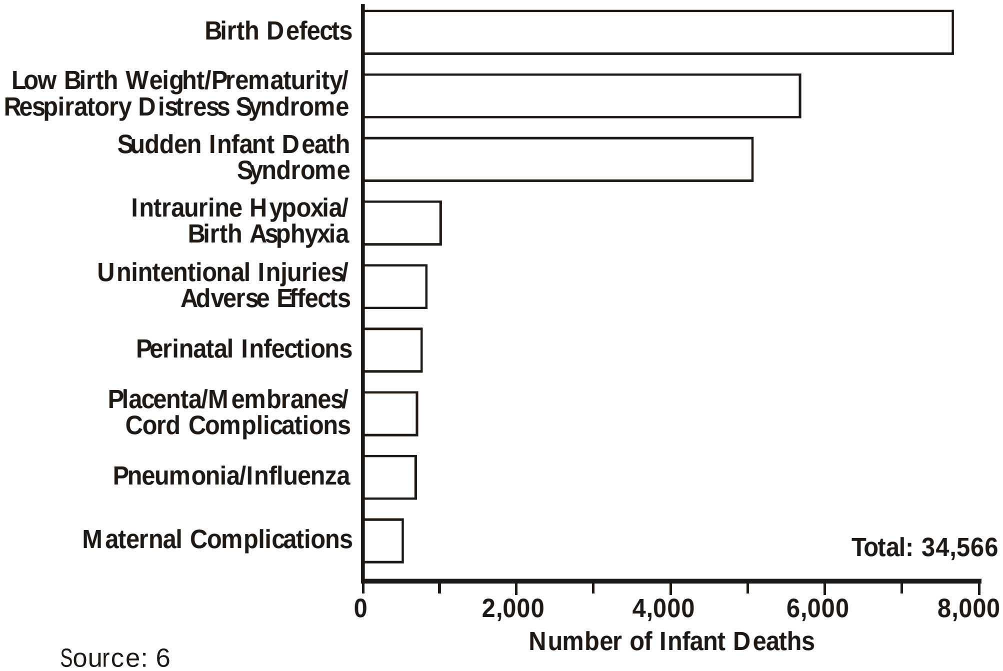
Bar Charts

The simplest bar chart is used to display the data from a one-variable table (see page 207). Each value or category of the variable is represented by a bar. The length of the bar is proportional to the number of persons or events in that category. Figure 4.17 shows the number of infant deaths by cause in the United States. This presentation of the data makes it very easy to compare the relative size of the different causes and to see that birth defects are the most common cause of infant mortality.

Variables shown in bar charts are either discrete and noncontinuous (e.g., race; sex) or are treated as though they were discrete and noncontinuous (e.g., age groups rather than age intervals along an axis).

Bars can be presented either horizontally or vertically. The length or height of each bar is proportional to the frequency of the event in that category. For this reason, **a scale break should not be used with a bar chart** since this could lead to misinterpretation in comparing the magnitude of different categories.

Figure 4.17
Example of horizontal bar chart:
Number of infant deaths by leading causes, United States, 1983



Source: 6

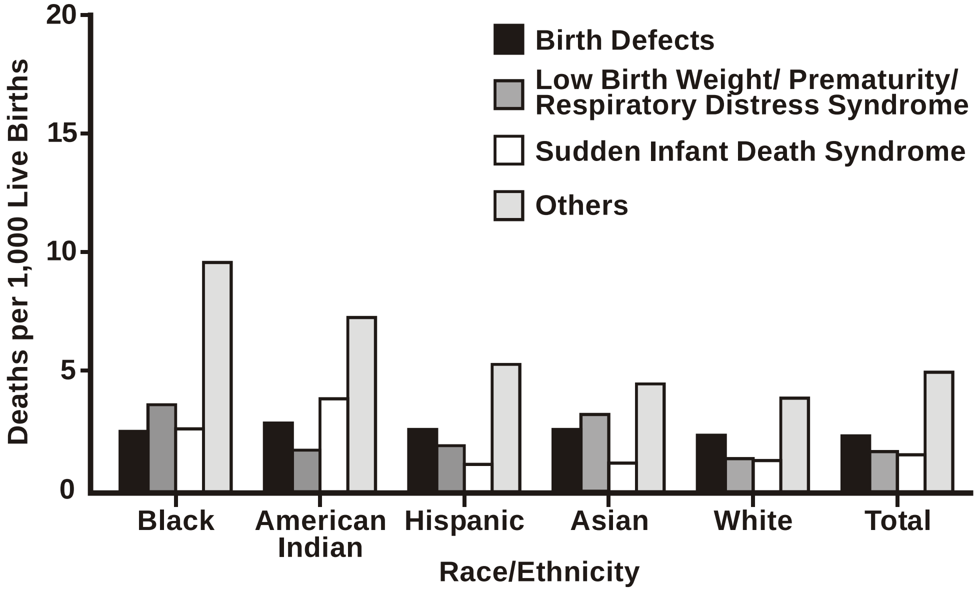
A vertical bar chart differs from a histogram in that the bars of a bar chart are separated while the bars of a histogram are joined. This distinction follows from the type of variable used on the *x-axis*. A histogram is used to show the frequency distribution of a continuous variable such as age or serum cholesterol or dates of onset during an epidemic. A bar chart is used to show the frequency distribution of a variable with discrete, noncontinuous categories such as sex or race or state.

Grouped Bar Charts

A grouped bar chart is used to illustrate data from two-variable or three-variable tables, when an outcome variable has only two categories. Bars within a group are usually adjoining. The bars must be illustrated distinctively and described in a legend. It is best to limit the number of bars within a group to no more than three. As you can see in Figure 4.18, it is difficult to interpret the data when the chart contains so many bars.

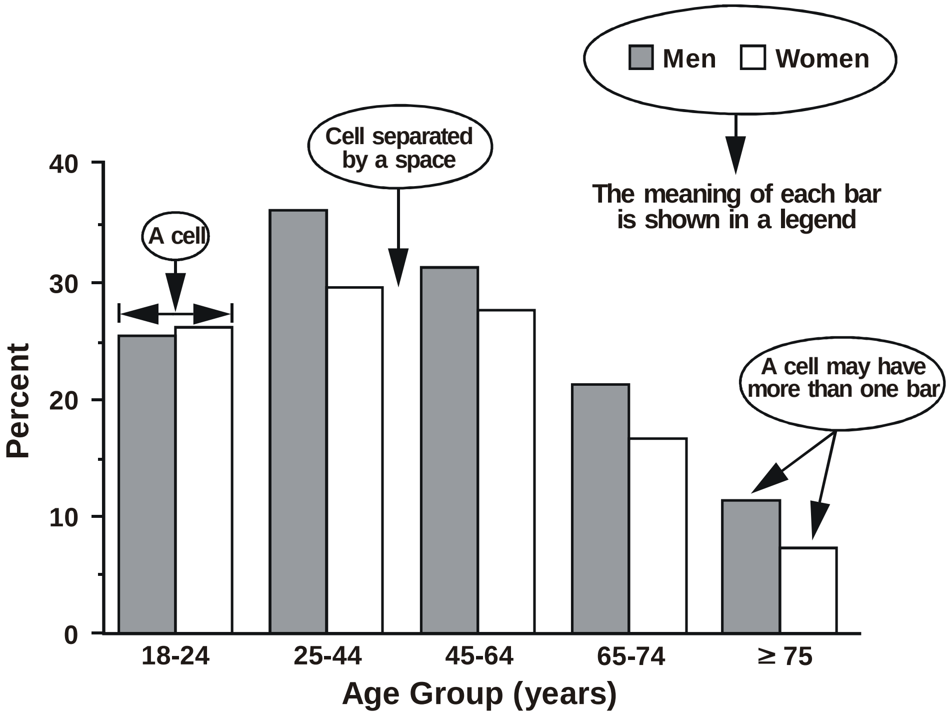
The bar chart in Figure 4.19 represents three variables: age, sex, and current smoking status. Current smoking status is the outcome variable and has two categories: yes or no. The bars represent the 10 age-sex categories. The height of each bar is proportional to the percentage of current smokers in each age-sex category.

Figure 4.18
Underlying cause of infant mortality among racial/ethnic groups, United States, 1983



Source 6

Figure 4.19
Example of vertical bar chart with annotation: Percentage of adults who were current cigarette smokers (persons ≥18 years of age who reported having smoked at least 100 cigarettes and who were currently smoking) by sex and age, United States, 1988



Source: 10

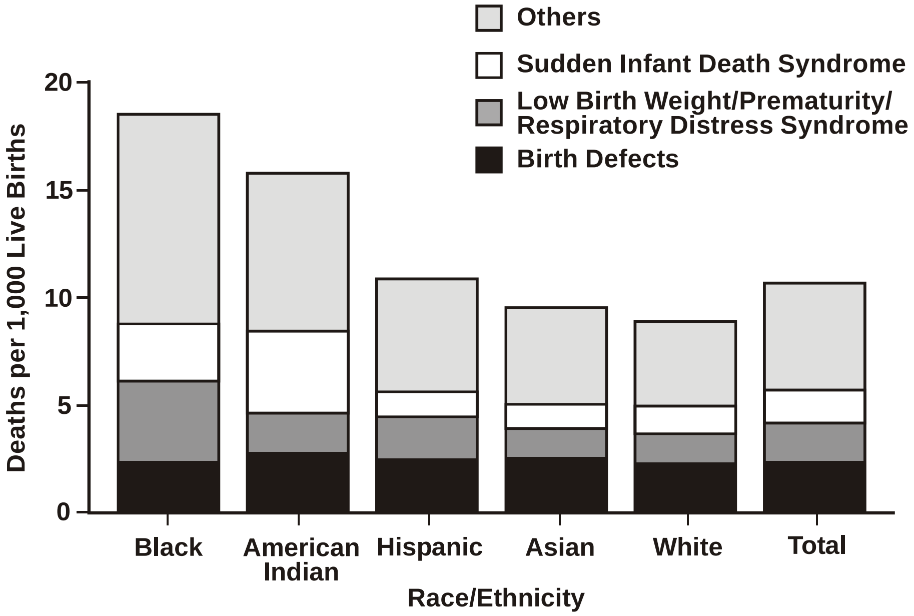
Stacked Bar Charts

You can also show categories of a second variable as components of the bars that represent the first variable, as in Figure 4.20. Notice that a stacked bar chart can be difficult to interpret because, except for the bottom component, the components do not rest on a flat baseline.

Deviation Bar Charts

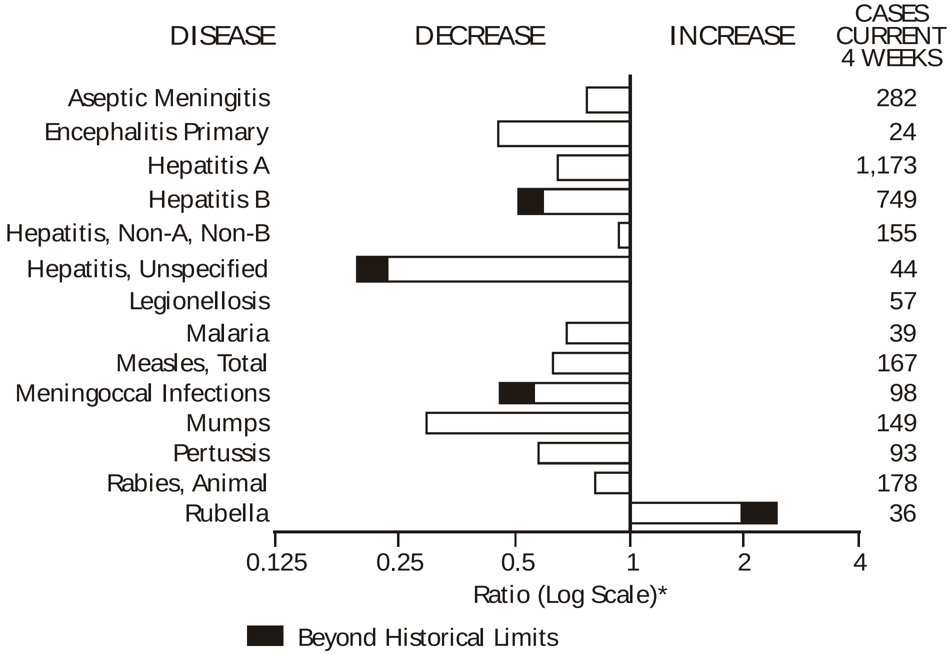
We can also use bar charts to show deviations in a variable, both positive and negative, from a baseline. Figure 4.21 shows such a deviation bar chart of selected reportable diseases in the United States. A similar chart appears weekly in CDC's *Morbidity and Mortality Weekly Report*. In this chart, the number of cases reported during the past 4 weeks are compared to the number reported during comparable periods of the past few years. The deviation to the right for rubella indicates an increase over historical levels. The deviations to the left indicate declines in reported cases compared to past levels. In this particular chart, the *x-axis* is on a logarithmic scale, so that a 50% reduction (one-half of the cases) and a doubling (50% increase) of cases would be represented by bars of the same length, though in opposite directions. Values beyond historical limits (comparable to 95% confidence limits) are highlighted for special attention.

Figure 4.20
Underlying causes of infant mortality among racial/ethnic groups, United States, 1983



Source 6

Figure 4.21
Notifiable Disease Reports, comparisons of 4-week totals ending January 26, 1991 with historical data, United States, 1991



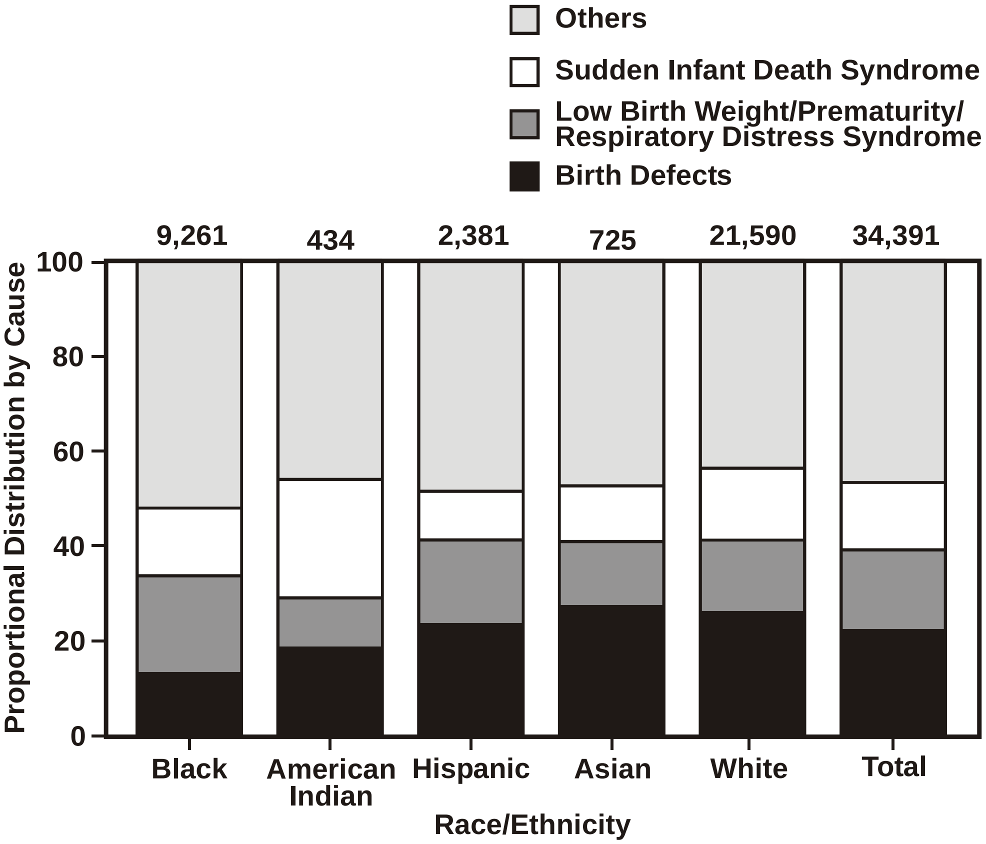
* Ratio of current 4-week total to mean of 15 4-week totals (from previous, comparable, and subsequent 4-week periods for the past 5 years). The point where the black area begins is based on the mean and two standard deviations of these 4-week totals.

Source 8

100% Component Bar Charts

In a variant of a stacked bar chart, we make all of the bars the same height (or length) and show the components as percents of the total rather than as actual values. This type of chart is useful for comparing the contribution of different components to each of the categories of the main variable. Figure 4.22 shows a 100% component bar chart. Notice that this type of bar chart is not useful for comparing the relative sizes of the various categories of the main variable (in this case, race/ethnicity); only the totals given above the bars indicate that the categories differed in size.

Figure 4.22
Underlying cause of infant mortality among racial/ethnic groups, United States, 1983



Source 6

How To Construct a Bar Chart

To construct a bar chart, observe the following guidelines:

- Arrange the categories that define the bars, or groups of bars, in a natural order, such as alphabetical or by increasing age, or in an order that will produce increasing or decreasing bar lengths
- Position the bars either vertically or horizontally as you prefer, except for deviation bar charts, in which the bars are usually positioned horizontally
- Make all of the bars the same width (which can be whatever looks in good proportion to you)
- Make the length of bars in proportion to the frequency of the event. Do not use a scale break, because it could lead to misinterpretation in comparing the size of different categories
- Show no more than three bars within a group of bars
- Leave a space between adjacent groups of bars, but not between bars within a group (see Figure 4.19)
- Code different variables by differences in bar color, shading, cross-hatching, etc. and include a legend that interprets your code

Exercise 4.6

Use the data in Table 4.12 to draw a stacked bar chart, a grouped bar chart, and a 100% component bar chart to illustrate the differences in the age distribution of syphilis cases among white males, white females, black males, and black females. What information is best conveyed by each chart? Graph paper is provided in Appendix D.

Table 4.12
Number of primary and secondary syphilis cases by age, sex, and race, United States, 1989

	White		Black		
Age group (years)	**Males**	**Females**	**Males**	**Females**	**Total**
<20	90	267	1,443	2,422	4,222
20-29	957	908	8,180	8,093	18,138
30-39	931	478	6,893	3,676	11,978
≥40	826	160	3,860	941	5,787
Total	**2,804**	**1,813**	**20,376**	**15,132**	**40,125**

Source: 12

Answer on pages 274-276.

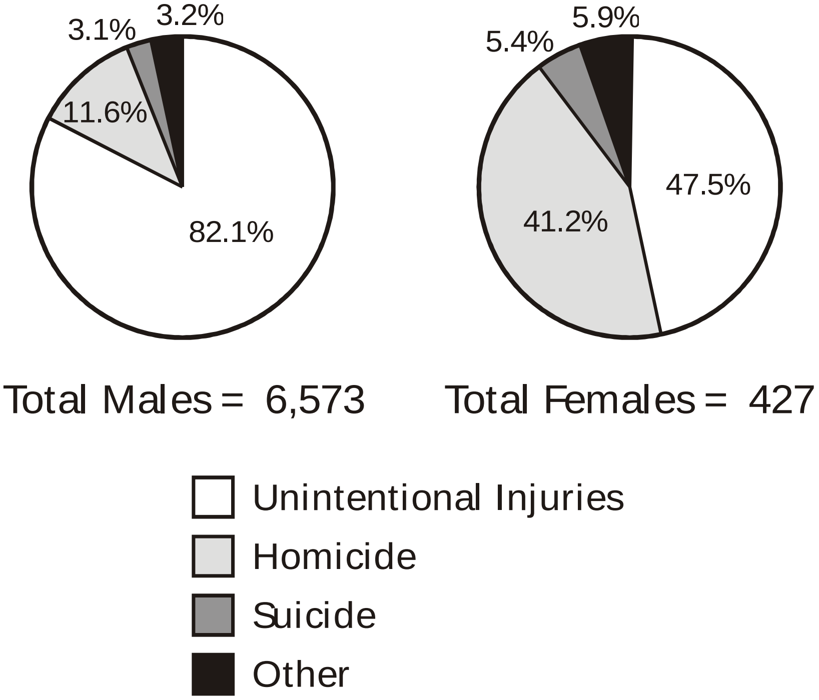
Pie Charts

A pie chart is a simple, easily understood chart in which the size of the “slices” show the proportional contribution of each component part. Pie charts are useful for showing the component parts of a single group or variable.

Graph paper is available commercially that has the circumference of a circle marked into 100 equal parts. This type of graph paper is called polar coordinate graph paper and an example is provided in Appendix D. Conventionally, you begin at 12 o’clock and arrange your component slices from largest to smallest, proceeding clockwise, although you may put the categories “other” and “unknown” last. You may use differences in shading to distinguish between slices. You should show somewhere on the graph what 100% represents, and because our eyes do not accurately gauge the area of the slices, you should indicate what percentage each slice represents either inside or near each slice.

Multiple pie charts as in Figure 4.23, are not optimal for comparing the same components in more than one group or variables, because it is difficult to compare components between two or more pie charts. When we want to compare the components of more than one group or variable, we use a 100% component bar chart.

Figure 4.23
Manner of traumatic deaths for male and female workers in the United States, 1980-1985

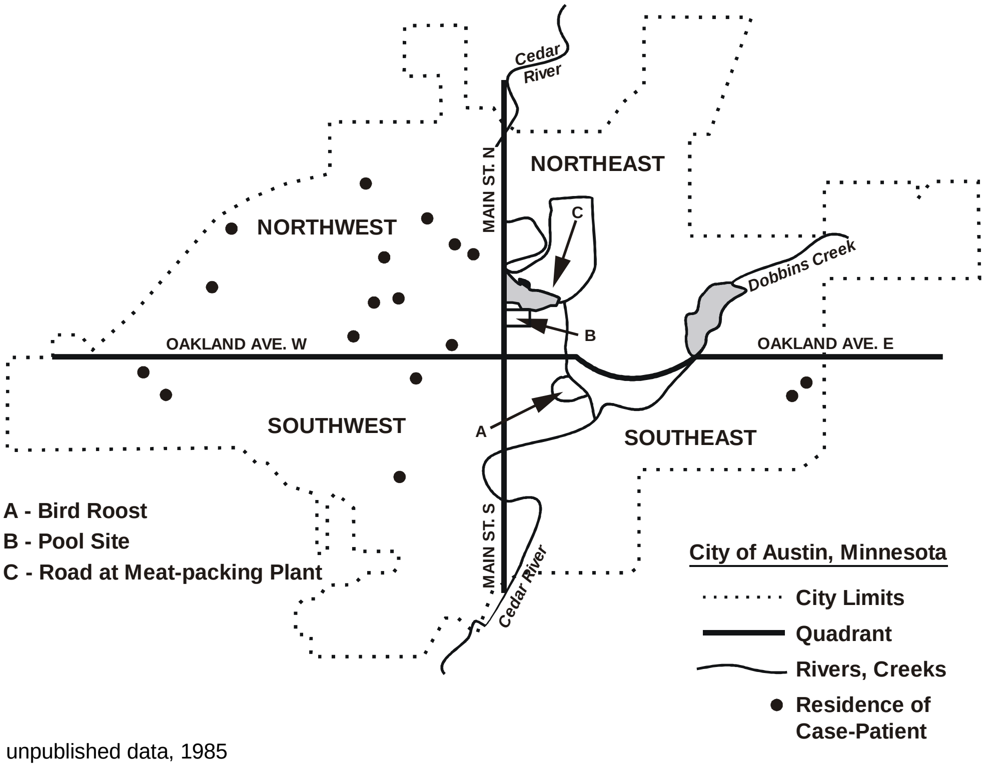


Source: 11

Maps (Geographic Coordinate Charts)

Maps or geographic coordinate charts are used to show the location of events or attributes. Spot maps and area maps are commonly used examples of this type of chart. Spot maps use dots or other symbols to show where an event took place or a disease condition exists. Figure 4.24 is an example of a spot map.

Figure 4.24
Example of spot map: Histoplasmosis by residence
Austin, Minnesota, October-November 1984



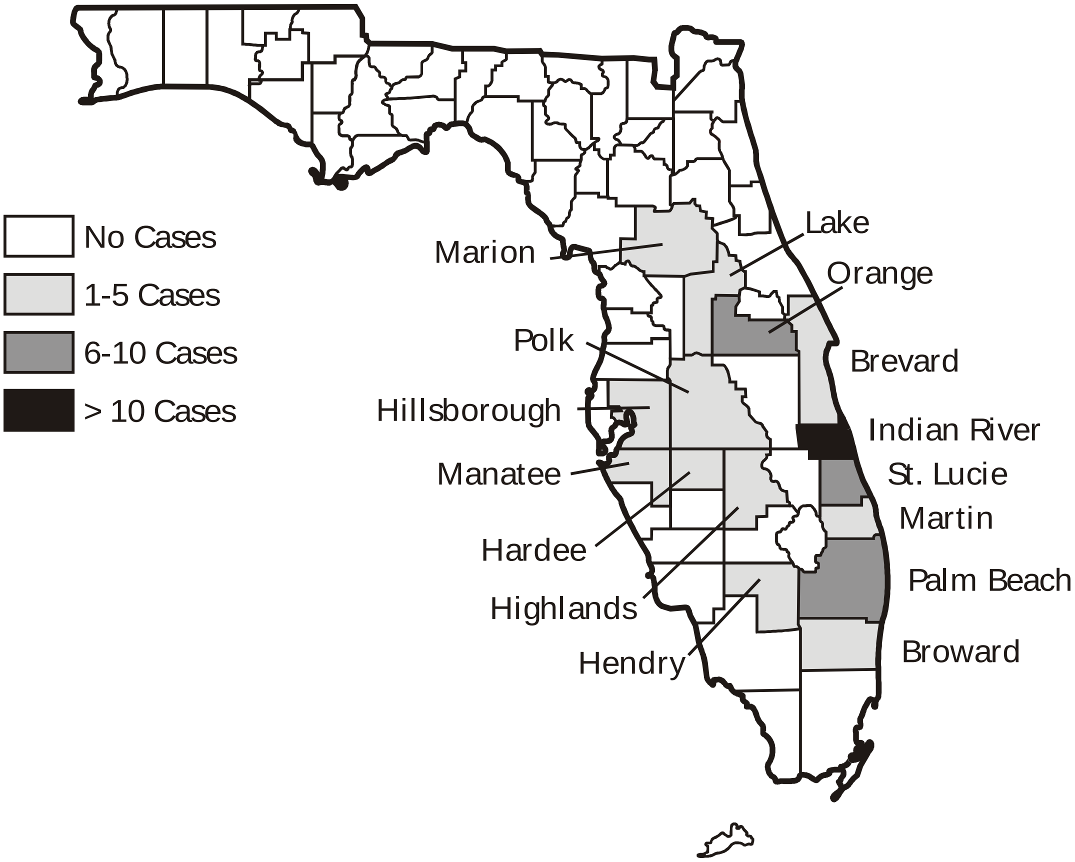
Source: CDC, unpublished data, 1985

To make a **spot map**, place a dot or other symbol on the map at the site where the event occurred or the condition exists. If events are clustered at one location, making it difficult to distinguish between dots, you can use coded symbols (e.g., • = 1 case, ■ = 2 cases, ▲ = 3 cases, etc.) that indicate the occurrence of more than one event.

A spot map is useful for showing the geographic distribution of an event, but—because it does not take into consideration the size of the population at risk—it does not show the **risk** of the event occurring in that particular place, for example, the risk of a resident acquiring a particular disease. Even when a spot map shows a large number of dots in the same area, the risk of acquiring the disease plotted may not be great there if that area is densely populated.

An area map uses shaded or coded areas to show either the incidence of an event in subareas, or the distribution of some condition over a geographic area. Figure 4.25 is an example of an area map.

Figure 4.25
Confirmed and presumptive cases of St. Louis encephalitis by county of residence, Florida, July–October 1990



Source: 7

We can show either numbers or rates with an **area map**. Figure 4.25 shows the numbers of cases of St. Louis encephalitis in different Florida counties in 1990. As with a spot map, this does not show the risk to persons living in these counties of acquiring St. Louis encephalitis. By showing rates in an area map, however, we can illustrate the differences in the risk of an event occurring in different areas. When we use rates, we must calculate a specific rate for each area—that is, we must divide the number of cases in each area by the population at risk in the same area.

Exercise 4.7

Using the cervical cancer mortality data in Table 4.9 on page 221, construct two area maps based on the first two strategies for categorizing data into four class intervals as described on pages 219-223. Maps of the United States are provided in Appendix D.

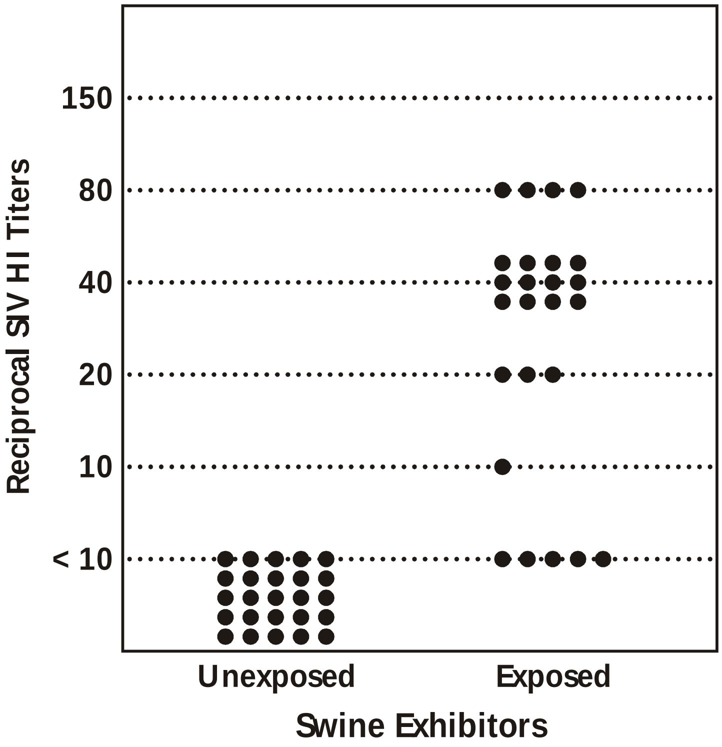
Answer on page 277.

Dot Plots and Box Plots

A **dot plot** is similar to a scatter diagram because it plots one variable against another. In a dot plot, however, the variable on the *x-axis* is not continuous—it represents discrete categories of a noncontinuous variable. As shown in Figure 4.26, we plot an observation by entering a dot over the appropriate *x* category at the level of the appropriate *y* value; and we show as many dots at that position as there are observations with those same values. Notice in Figure 4.26 that the different vertical positions of the 12 dots at the intersection of "Exposed" and "40" do not indicate their titer levels: they all have titer levels of 40. The dots are placed on different lines to facilitate showing them as a unit. Similarly, the 25 dots at "Unexposed" all represent a titer level of <10.

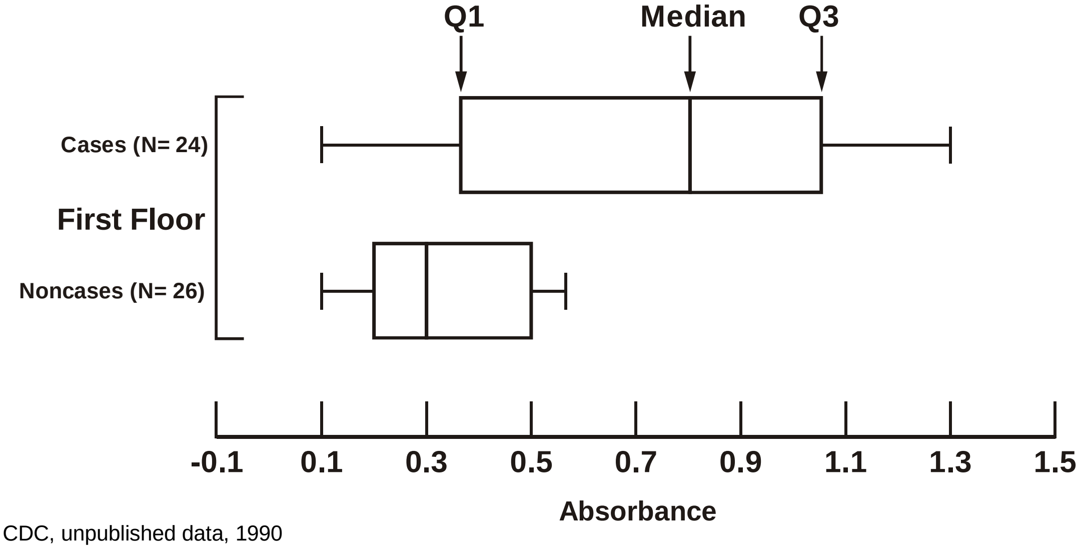
We use a dot plot to make a visual comparison of the actual data points of two noncontinuous variables. If we instead want to compare the *distributions* of noncontinuous variables, we use a **box plot**. In a box plot, we show the distributions of data as "box and whiskers" diagrams, shown in Figure 4.27. The "box" represents the middle 50% or interquartile range of the data, and the "whiskers" extend to the minimum and maximum values. We mark the position of the median with a vertical line inside the box. Thus, with a box plot we can show (and compare) the central location (median), dispersion (interquartile range and range), and any tendency toward skewness, which is indicated if the median line is not centered in the box.

Figure 4.26
Example of dot plot: Results of swine influenza virus (SIV) hemagglutination-inhibition (HI) antibody testing among exposed and unexposed swine exhibitors, Wisconsin, 1988



Source: 26

Figure 4.27
Example of box plot: Results of indirect ELISA for IgG antibodies to parainfluenza type I virus in convalescent phase serum specimens from cases to noncases, Baltimore County, Maryland, January 1990



Source: CDC, unpublished data, 1990

A Comment About Using Computer Technology

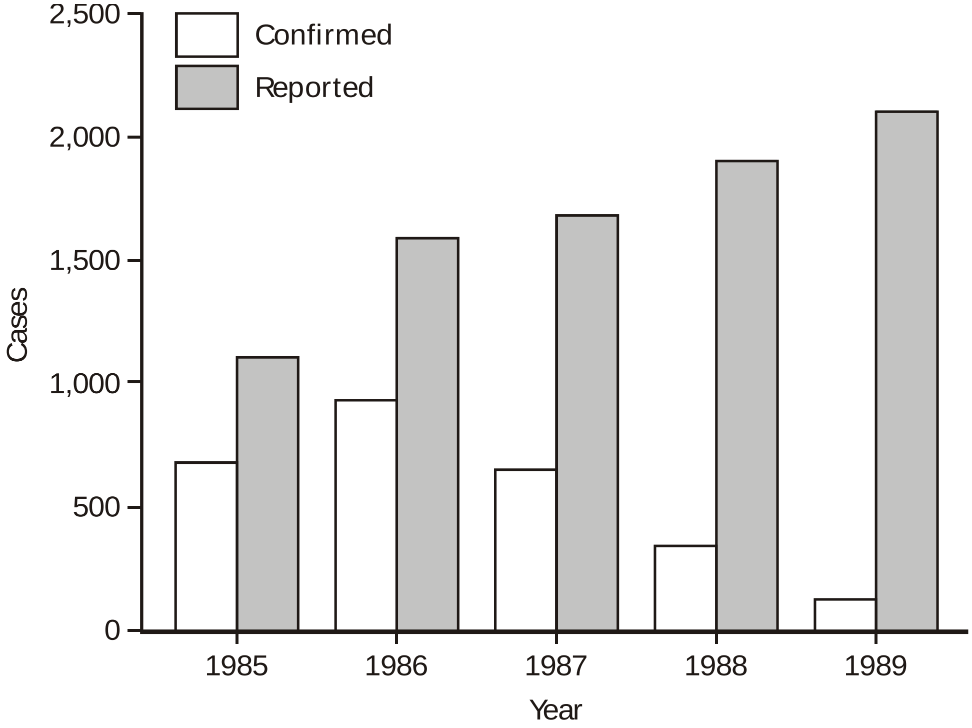
A large number of software packages for the personal computer are available that can help us make tables, graphs, and charts. Most of these packages are quite useful, particularly in letting us redraw a graph with only a few keystrokes. With these packages, finding the best epidemic curve is no longer an onerous and tedious task: We can now quickly and easily draw a number of curves with different class intervals on the *x-axis*.

On the other hand, we are sometimes tempted to let the software dictate the graph. For example, many packages can draw bar charts and pie charts that appear three-dimensional. Does that mean we should develop three-dimensional charts? We need to keep our purpose in mind: to communicate information to others. Will three-dimensional charts communicate the information better than a two-dimensional chart?

Decide for yourself: Does the three-dimensional chart in Figure 4.28b provide any more information than the two-dimensional bar chart in Figure 4.28a? Which is easier to interpret?

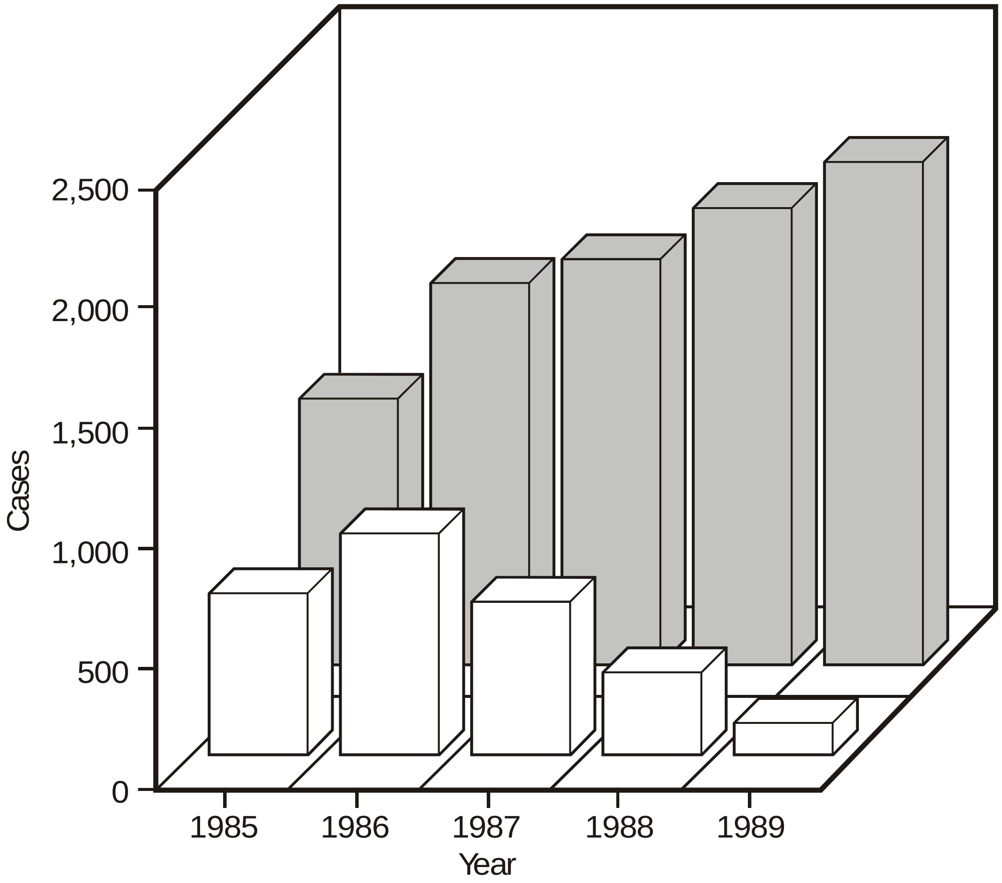
If we wanted to focus on the trends over time for confirmed and for reported cases, perhaps the three-dimensional chart is preferable. However, an arithmetic-scale line graph with two lines might be best of all. A problem common to three-dimensional bar charts is that a bar in the front

Figure 4.28a
Example of two-dimensional bar chart:
Reported and confirmed polio cases by year, the Americas, 1985-1989



Source: 5

Figure 4.28b
Example of three-dimensional bar chart:
Reported and confirmed polio cases by year, the Americas, 1985-1989



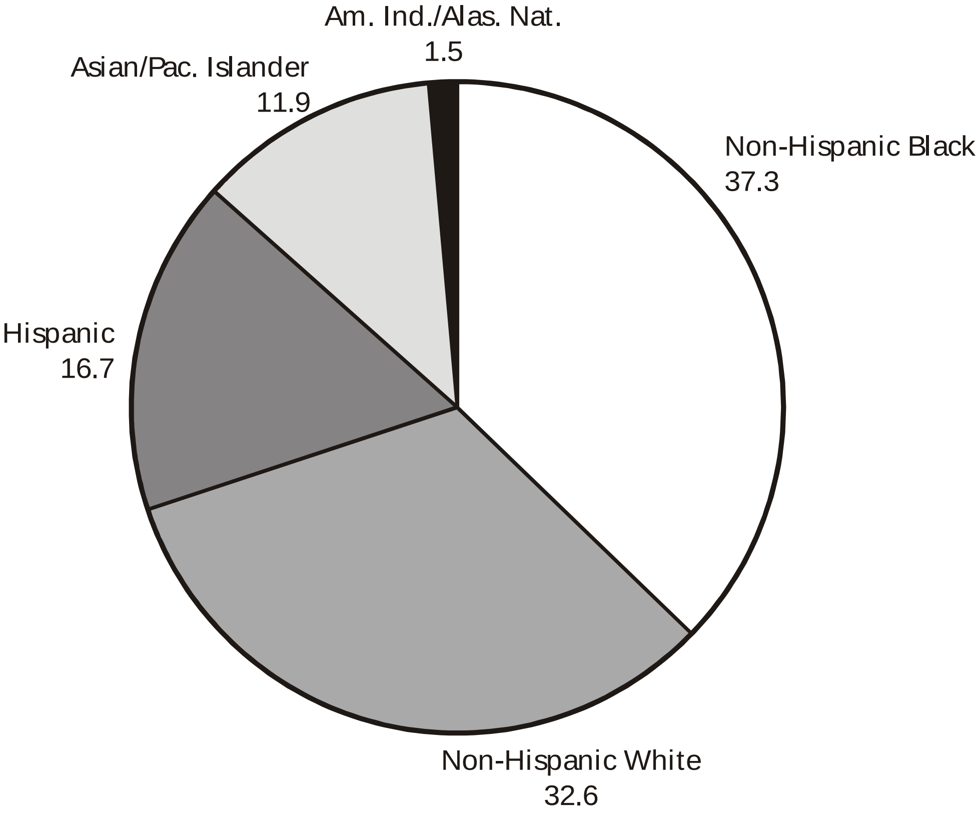
Source: 5

row may block a bar in the back row. Suppose that we are interested in the ratio of confirmed to reported cases each year. We see immediately from the two-dimensional bar chart that the number of confirmed cases in 1985 is approximately two-thirds of the number of reported cases in 1985. How long do you have to look at the three-dimensional chart to reach that same conclusion? Now compare that ratio of confirmed to reported cases for all five years. If you need to communicate this information with a slide in 20 seconds during a 10-minute oral presentation, which figure would you show?

Does the three-dimensional pie chart in Figure 4.29b provide any more information than the two-dimensional chart in Figure 4.29a? Can you judge the relative sizes of the components as well in the three-dimensional version? Look at the three-dimensional pie and block out the percentages for Hispanics and Asian/Pacific Islanders. Can you really tell which wedge is bigger and by how much? We think you can’t. Can you tell from the two-dimensional pie? Remember that size is the whole purpose of a pie chart.

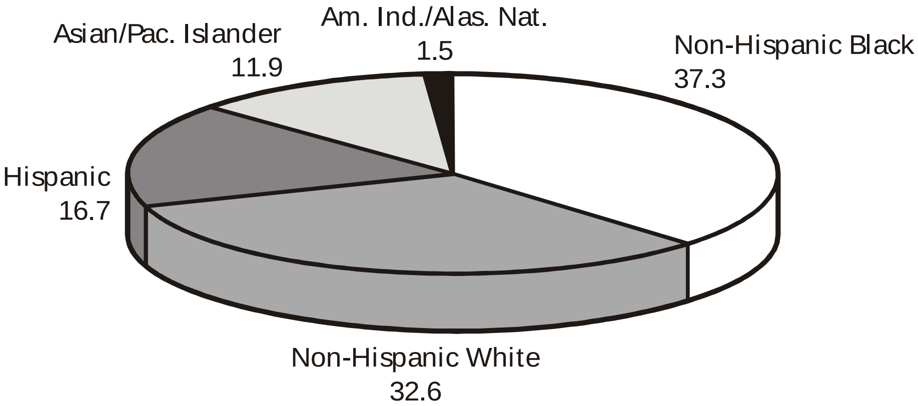
The addition of gimmicky features to a figure which adds no information and which may even promote misinterpretation has been termed **chartjunk** (25).

Figure 4.29a
Example of two-dimensional pie chart:
Percentage of tuberculosis cases by race and ethnicity,
United States, 1989 (n = 23,495)



Source: 12

Figure 4.29b
Example of three-dimensional pie chart:
Percentage of tuberculosis cases by race and ethnicity,
United States, 1989 (n = 23,495)



Source: 12

Many people misuse technology in selecting color, particularly for slides that accompany oral presentations. If you use colors at all, follow these recommendations:

- Select colors so that all components of the graph—title, axes, data plots, legends—stand out clearly from the background, and so that each plotted series of data can be distinguished from the others.
- Avoid contrasting red and green, because up to 10% of males in the audience may have some degree of color blindness.
- When possible, select colors so that they communicate information. For example, consider an area map in which states are divided into four groups according to their rates for a particular disease. Rather than choosing colors solely for appearance, you might use a light color or shade for the states with the lowest rates and progressively darker colors or shades for the groups with progressively higher rates. In this way, the colors contribute to, rather than distort or distract from the information you want to convey.

Finally, with some software packages you cannot produce some of the types of graphs covered in this manual. In particular, some software packages cannot create a histogram; instead they produce a bar chart. Your graphs should be dictated by your data and the relationships you want to communicate visually, not by the technology at hand. If the software you have cannot accommodate your data, don’t compromise the integrity of the data or its presentation. Use different software!

Selecting and Constructing Tables, Graphs, and Charts

To convey the messages of epidemiologic findings, you must first select the best illustration method. But even the best method must be constructed properly or the message will be lost. The tables in this section provide guidance in the selection of illustration methods and construction of tables, graphs, and charts.

Table 4.13
Guide to selecting a graph or chart to illustrate epidemiologic data

Type of Graph or Chart	When to Use
Arithmetic-scale line graph	Trends in numbers or rates over time
Semilogarithmic-scale line graph	1. Emphasize rate of change over time 2. Display values ranging over more than 2 orders of magnitude
Histogram	1. Frequency distribution of continuous variable 2. Number of cases during epidemic (epidemic curve) or over time
Frequency polygon	Frequency distribution of continuous variable, especially to show components
Cumulative frequency	Cumulative frequency for continuous variables
Scatter diagram	Plot association between two variables
Simple bar chart	Compare size or frequency of different categories of a single variable
Grouped bar chart	Compare size or frequency of different categories of 2–4 series of data
Stacked bar chart	Compare totals and illustrate component parts of the total among different groups
Deviation bar chart	Illustrate differences, both positive and negative, from baseline
100% component bar chart	Compare how components contribute to the whole in different groups
Pie chart	Show components of a whole
Spot map	Show location of cases or events
Area map	Display events or rates geographically
Box plot	Visualize statistical characteristics (median, range, skewness) of a variable

Table 4.14
Selecting a method of illustrating epidemiologic data

If Data Are:		And These Conditions Apply:		Then Choose:
Time series		Numbers of cases (epidemic or secular trend)	1 or 2 sets	Histogram
			2 or more sets	Frequency polygon
		Rates	Range of values ≤ 2 orders of magnitude	Arithmetic scale line graph
			Range of values ≤ 2 orders of magnitude	Semilogarithmic scale line graph
Continuous data other than time series		Frequency distribution		Histogram or frequency polygon
Data with discrete categories (other than place)				Bar chart or pie chart
Place	Number of cases	Not readily identified on map		Bar chart
		Readily identified on map	Specific site important	Spot map
			Specific site unimportant	Area map
	Rates			Area map

Table 4.15
Checklist for construction of tables, graphs, charts, and visuals

Checklist for Tables

1. Title
   - Does the table have a title?
   - Does the title describe the content, including subject, person, place, and time?
   - Is the title preceded by the designation “Table #”? (“Table” is used for typed text; “Figure” for graphs, charts, and maps. Separate numerical sequences are used for tables and figures in the same document [e.g., Table 1, Table 2, Figure 1, Figure 2]).
2. Rows and columns
   - Is each row and each column labeled clearly and concisely?
   - Are the specific units of measurement shown? (e.g., years, mm Hg, mg/dl, rate per 100,000, etc.).
   - Are the categories appropriate for the data?
   - Are the row and column totals provided?
3. Footnotes
   - Are all codes, abbreviations, or symbols explained?
   - Are all exclusions noted?
   - If the data are not original, is the source provided?

Checklist for Graphs and Charts

1. Title
   - Does the graph or chart have a title?
   - Does the title describe the content, including subject, person, place, and time?
   - Is the title preceded by the designation “Figure #”? (“Table” is used for typed text; “Figure” for graphs, charts, and maps. Separate numerical sequences are used for tables and figures in the same document [e.g., Table 1, Table 2, Figure 1, Figure 2]).
2. Axes
   - Is each axis labeled clearly and concisely?
   - Are the specific units of measurement included as part of the label? (e.g., years, mm Hg, mg/dl, rate per 100,000, etc.)
   - Are the scale divisions on the axes clearly indicated?
   - Are the scales for each axis appropriate for the data?
   - Does the *y*-axis start at zero?
   - If a scale break is used with a scale line graph, is it clearly identified?
   - Has a scale break been used with a histogram, frequency polygon, or bar chart? (Answer should be **NO!**)
   - Are the axes drawn heavier than the other coordinate lines?
3. Coordinate lines
   - Does the figure include only as many coordinate lines as are necessary to guide the eye? (Often, these are unnecessary.)

Table 4.15
Checklist for construction of tables, graphs, charts, and visuals – continued

4. Data plots
   - Are the plots drawn clearly?
   - If more than one series of data or components are shown, are they clearly distinguishable on the graph?
   - Is each series or component labeled on the graph, or in a legend or key?
   - If color or shading is used on an area map, does an increase in color or shading correspond to an increase in the variable being shown?
5. Footnotes
   - Are all codes, abbreviations, or symbols explained?
   - Are all exclusions noted?
   - If the data are not original, is the source provided?
6. Visual Display
   - Does the figure include any information that is not necessary?
   - Is the figure positioned on the page for optimal readability?
   - Do font sizes and colors improve readability?

Checklist for Effective Visuals (14)

1. Legibility (make sure your audience can easily read your visuals)
   - Can your overhead transparencies be read easily from 6 feet when not projected?
   - Can your 35mm slides be read easily from 1 foot when not projected?
   - When projected, can your visuals be read from the farthest parts of the room?
2. Simplicity (keep the message simple)
   - Have you used plain words?
   - Is the information presented in the language of the audience?
   - Have you used only “key” words?
   - Have you omitted conjunctions, prepositions, etc.?
   - Is each visual limited to only one major idea/concept/theme?
   - Does each visual have no more than 3 colors?
   - Are there no more than 35 letters and numbers on each visual?
   - Are there no more than 6 lines of narration and 6 words per line?

Table 4.15
Checklist for construction of tables, graphs, charts, and visuals — continued

3. Colorfulness
   - The colors you select for your visuals will have an impact on the effect of your visuals. You should use warm/hot colors to emphasize, to highlight, to focus, or to reinforce key concepts. You should use cool/cold colors for background or to separate items. Use the table below to select the appropriate color for the effect you desire.

	Hot	**Warm**	**Cool**	**Cold**
Colors:	Reds	Light orange	Light blue	Dark blue
	Bright orange	Light yellow	Light green	Dark green
	Bright yellow	Light gold	Light purple	Dark purple
	Bright gold	Browns	Light gray	Dark gray
Effect:	Exciting	Mild	Subdued	Somber

   - Are you using the best color combinations? The most important item should be in the most important color and have the greatest contrast with its background. The most legible color combinations are:

     Black on Yellow
     Black on White
     Dark Green on White
     Dark Blue on White
     White on Dark Blue

4. Accuracy
   Visuals become distractions when mistakes are spotted. Have someone who has not seen the visual before check for typos, inaccuracies, and errors in general.

5. Durability
   Transparencies and 35mm slides are the most durable of the visual aids. However, both require some protection from scratches. A clear sheet of acetate or Mylar will protect a transparency. Keep 35mm slides in a cool, dark place. If left in the light, colors will fade.

Summary

Tables, graphs, and charts are effective tools for summarizing and communicating data. Tables are commonly used to display numbers, rates, proportions, and cumulative percents. Because tables are intended to communicate information, most tables should have no more than two variables and no more than eight categories (class intervals) of any variable. Tables are sometimes used out of context by others, so they should be properly titled, labeled, and referenced.

Tables can be used with either nominal or continuous ordinal data. Nominal variables such as sex and state of residence have obvious categories. Continuous variables do not; class intervals must be created. For some diseases, standard class intervals for age have been adopted. Otherwise a variety of methods are available for establishing reasonable class intervals. These include class intervals with an equal number of people or observations in each; class intervals with a constant width; and class intervals based on the mean and standard deviation.

Graphs and charts are even more effective tools for communicating data rapidly. Although some people use the terms *graph* and *chart* interchangeably, in this Lesson *graph* refers to a figure with two coordinates, a horizontal *x-axis* and a vertical *y-axis*. In other words, both variables are continuous. For example, the *y-axis* commonly features number of cases or rate of disease; the *x-axis* usually represents time. In contrast, a *chart* refers to a figure with one continuous and one nominal variable. For example, the chart may feature number of cases (a continuous variable) by sex (a nominal variable).

Arithmetic-scale line graphs have traditionally been used to show trends in disease **rates** over time. Semilogarithmic-scale line graphs are preferred when the disease rates vary over two or more orders of magnitude. Histograms and frequency polygons are used to display frequency distributions. A special type of histogram known as an epidemic curve shows the **number** of cases by time of onset of illness or time of diagnosis during an epidemic period. The cases may be represented by squares which are stacked to form the columns of the histogram; the squares may be shaded to distinguish important characteristics of cases, such as fatal outcome.

Simple bar charts and pie charts are used to display the frequency distribution of a single variable. Grouped and stacked bar charts can display two or even three variables.

Spot maps pinpoint the location of each case or event. An area map uses shading or coloring to show different levels of disease numbers or rates in different areas.

When using these tools, it is important to remember their purpose: to summarize and to communicate. Glitzy and colorful are not necessarily better; sometimes less is more!

Answers to Exercises

Answer—Exercise 4.1 (page 212)

A.

**Occurrence of diarrhea by menu,
residents of Nursing Home A, 1989**

	Diarrhea status		
Menu	**Yes**	**No**	**Total**
A	12	5	17
B	0	7	7
C	0	4	4
D	2	4	6
E	0	1	1
F	0	1	1
Total	**14**	**22**	**36**

B.

**Occurrence of diarrhea by exposure to menu A,
residents of Nursing Home A, 1989**

		Diarrhea		
		Yes	**No**	**Total**
	Yes	12	5	17
Menu A	**No**	2	17	19
	Total	14	22	36

Answer—Exercise 4.2 (page 225)

Strategy 1: Divide the data into groups of similar size

Divide the list into three equal-sized groups of states:

50 states ÷ 3 = 16.67 states per group. Thus, two groups will contain 17 states and one group will contain 16 states.

Oklahoma (#17) could go in either group 1 or group 2, but since it has the same rate as Indiana (#16), it makes sense to put Oklahoma in group 1. Similarly, since Michigan (#34) could go in either group 2 or group 3 but has the same rate as Oregon (#33), Michigan should go in group 2.

Final categories:

States	Range of rates per 100,000	Number of states
1. OK-SC	4.1-5.6	17
2. MI-IL	3.3-4.0	17
3. UT-CA	1.8-3.2	16

Strategy 2: Base categories on the mean and standard deviation

Create 3 categories based on mean (3.70) and standard deviation (0.96):

upper limit of category 1 = mean – 1 standard deviation = 3.70 – 0.96 = 2.74

upper limit of category 2 = mean + 1 standard deviation = 3.70 + 0.96 = 4.66

upper limit of category 3 = maximum value = 5.6

Final Categories:

States	Range of rates per 100,000	Number of states
1. MS-SC	4.67-5.60	9
2. RI-NC	2.75-4.66	34
3. UT-WI	1.80-2.74	7

Strategy 3: Divide the range into equal class intervals

Divide the range by 3: (5.60 – 1.80) ÷ 3 = 1.267

Use multiples of 1.27 to create three categories, starting with 1.8:

1. 1.80 through (1.80 + 1.27) = 1.80 through 3.07
2. 3.08 through (1.80 + 2 × 1.27) = 3.08 through 4.34
3. 4.35 through (1.80 + 3 × 1.27) = 4.35 through 5.61

Final categories:

States	Range of rates per 100,000	Number of states
1. ME-SC	4.35-5.61	12
2. AZ-VT	3.08-4.34	25
3. UT-MA	1.80-3.07	13

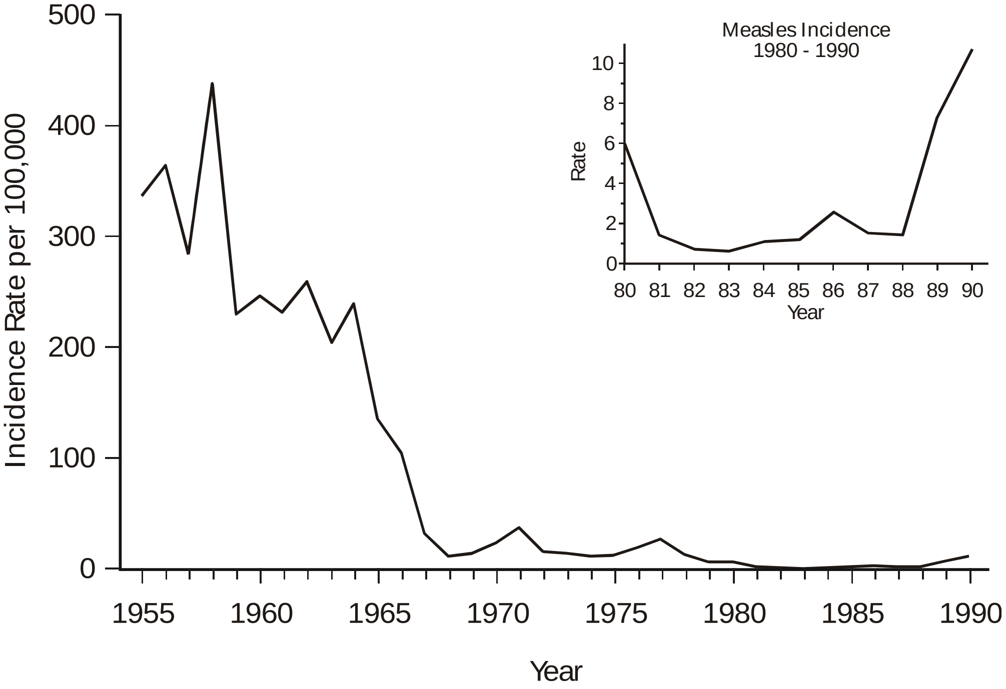
Or rounding categories:

States	Range of rates per 100,000	Number of states
1. ME-SC	4.4-5.6	12
2. AZ-VT	3.1-4.3	25
3. UT-MA	1.8-3.0	13

Answer—Exercise 4.3 (page 231)

A. and B.

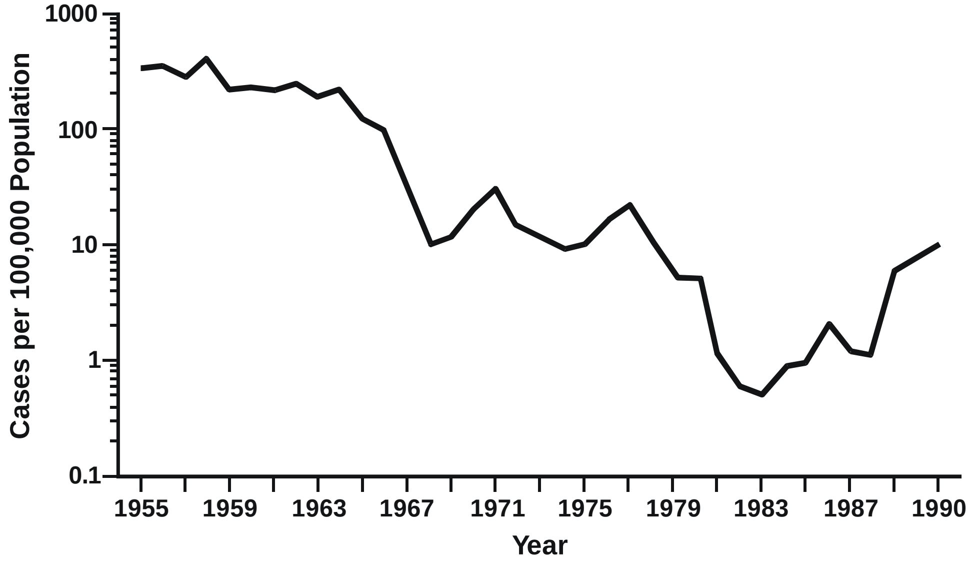
Figure 4.30
Annual measles incidence rates per 100,000, United States, 1955-1990; with inset of 1980-1990



Source: 12

Answer—Exercise 4.4 (page 235)

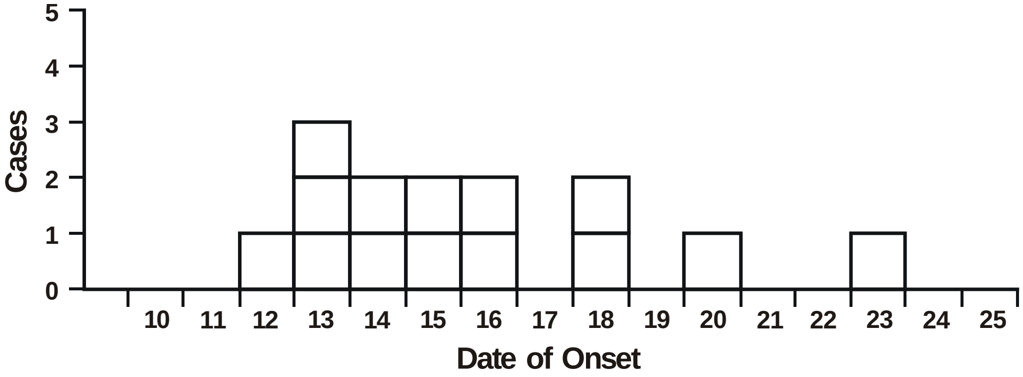
Figure 4.31
Annual measles incidence rates per 100,000, United States, 1955-1990



Source: 12

Answer—Exercise 4.5 (page 240)

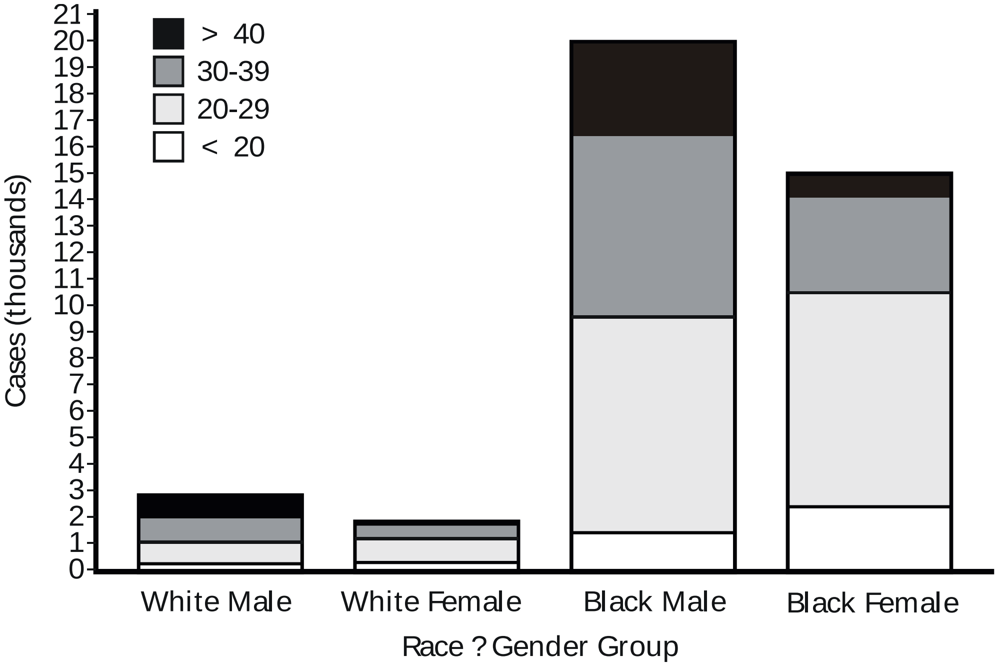
Figure 4.32
Outbreak of diarrheal disease in Nursing Home A, January 1989



This outbreak appeared to last just under two weeks, from January 12 to January 23. After the initial case on January 12th, the peak occurred the following day, with three cases on January 13. The curve was relatively flat after that, with two cases each on four of the next five days. Single cases occurred in January 20 and January 23.

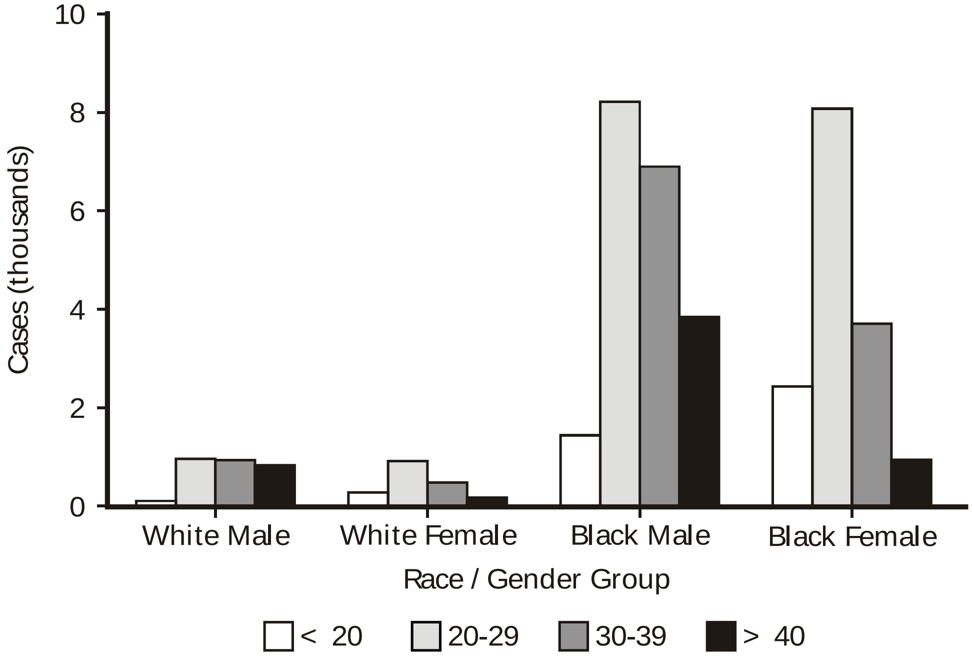
Answer—Exercise 4.6 (page 252)

Figure 4.33a
Stacked bar chart: Number of primary and secondary syphilis cases by age, sex, and race, 1989



Answer—Exercise 4.6 (continued)

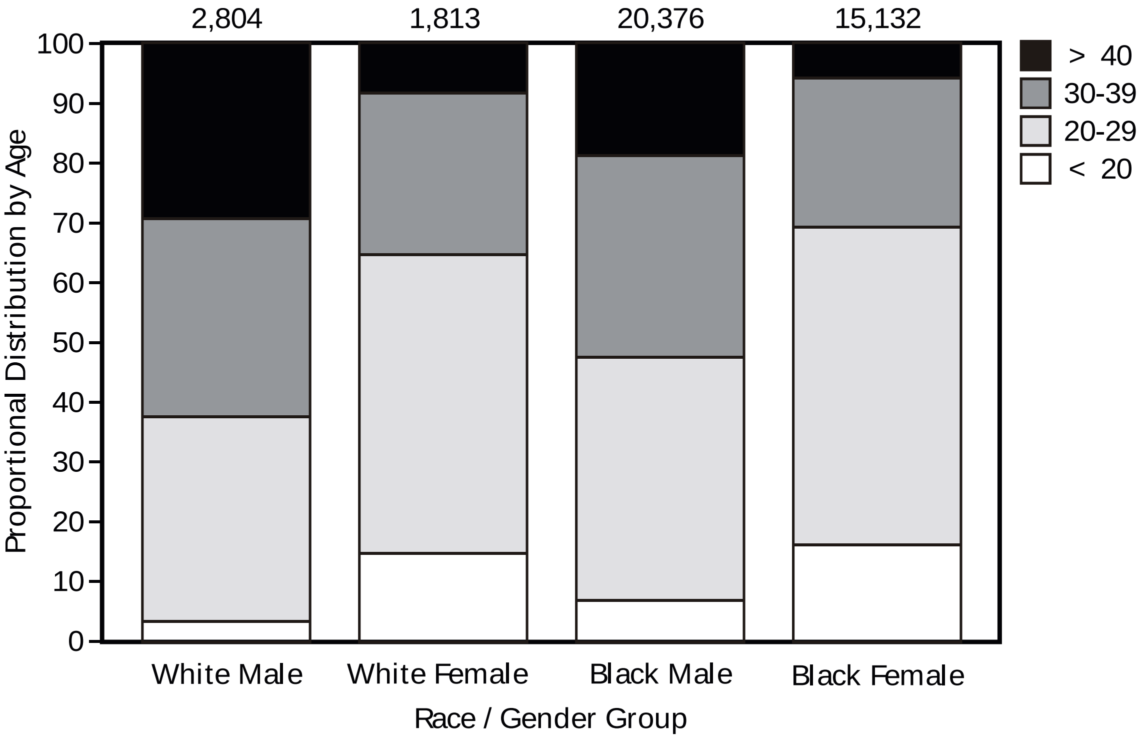
Figure 4.33b
Grouped bar chart: Number of primary and secondary syphilis cases by age, sex, and race, 1989



Source: 12

Answer—Exercise 4.6 (continued)

Figure 4.33c
100% component bar chart: Number of primary and secondary syphilis cases by age, sex, and race, 1989

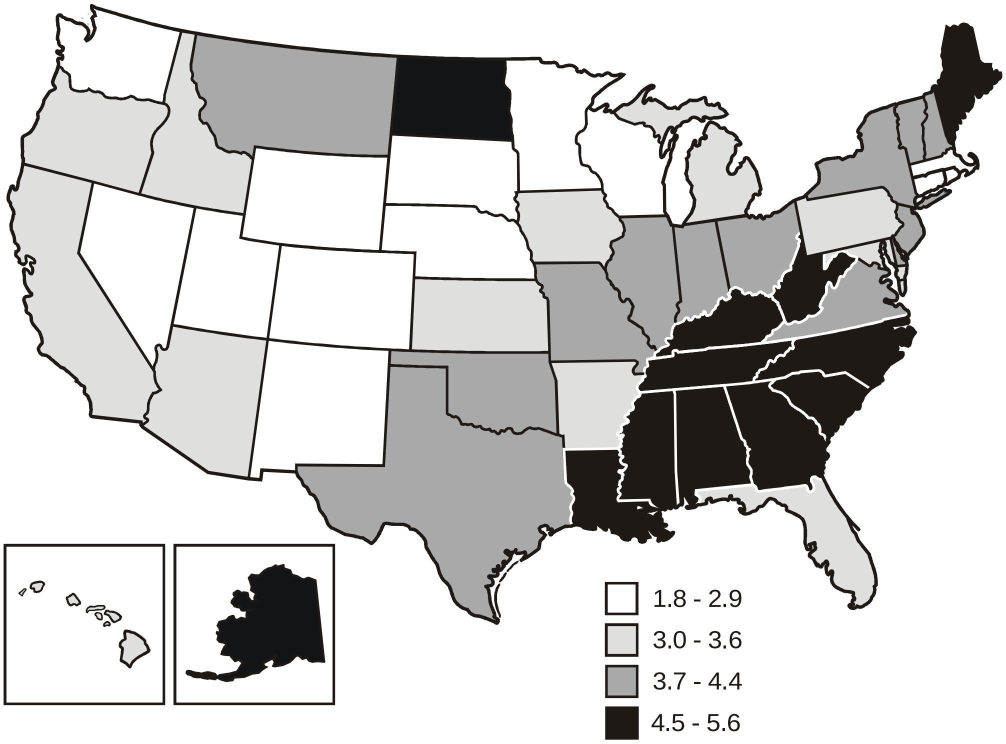


Source: 12

Answer—Exercise 4.7 (page 256)

A.

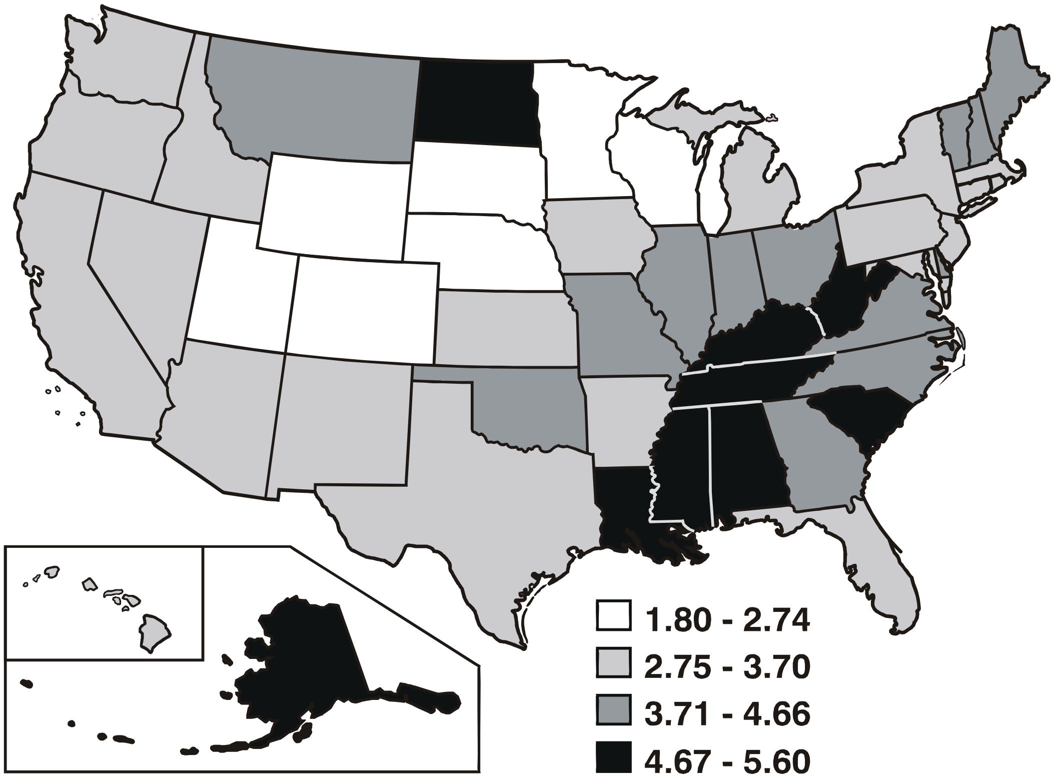
Figure 4.34a
Strategy 1: Mean annual age-adjusted cervical cancer mortality rates per 100,000 population by state, United States, 1984-1986



Source: 2

B.

Figure 4.34b
Strategy 2: Mean annual age-adjusted cervical cancer mortality rates per 100,000 population by state, United States, 1984-1986



Source: 2

Self-Assessment Quiz 4

Now that you have read Lesson 4 and have completed the exercises, you should be ready to take the self-assessment quiz. This quiz is designed to help you assess how well you have learned the content of this lesson. You may refer to the lesson text whenever you are unsure of the answer, but keep in mind that the final is a closed book examination. Circle ALL correct choices in each question.

1. Tables, graphs, and charts are important tools for which tasks of an epidemiologist? (Circle ALL that apply.)

   A. Data collection

   B. Data summarization (descriptive epidemiology)

   C. Data analysis

   D. Data presentation

2. Which two-by-two table is properly labeled?

   A.

	ILL	WELL	TOTAL
Exposed	a	c	H1
Unexposed	b	d	H2
Total	V1	V2	T

   B.

	ILL	WELL	TOTAL
Exposed	a	b	V1
Unexposed	c	d	V2
	H1	H2	T

   C.

	ILL	WELL	TOTAL
Exposed	a	b	H1
Unexposed	c	d	H2
Total	V1	V2	T

   D.

	Exposed	Unexposed	TOTAL
ILL	a	c	H1
WELL	b	d	H2
	V1	V2	T

Primary and secondary syphilis morbidity by age, United States, 1989

Age group (years)	Cases		
	Number	Percent	Cumulative Percent
≤14	230	0.5%	0.5%
15-19	4,378	9.9%	10.4%
20-24	10,405	23.6%	34.0%
25-29	9,610	21.8%	55.9%
30-34	8,648	19.6%	75.5%
35-44	6,901	15.7%	91.2%
45-54	2,631	6.0%	97.2%
55+	1,278	2.9%	100.1%
Total	**44,081**	**100.0%***	**100.0%**

*Percentages do not add to 100.0% due to rounding.

3. The table shown above is an example of a/an:

A. one-variable table

B. two-variable table

C. three-variable table

D. four-variable table

4. The maximum number of variables that should be cross-tabulated in a single table is:

A. 1

B. 2

C. 3

D. 4

5. The best time to create table shells is:

A. just before planning the study

B. as part of planning the study

C. just after collecting the data

D. just before analyzing the data

E. as part of analyzing the data

6. Recommended methods for creating categories for continuous variables include: (Circle ALL that apply)

A. basing the categories on the mean and standard deviation

B. dividing the data into categories with similar numbers of observations in each

C. dividing the range into equal class intervals

D. using the categories which are considered standard for the condition

E. using the same categories as your population data are grouped

7. The Lesson illustrates three strategies for creating class intervals for continuous variables. Which of the following sets of class intervals shown in the answer list (A-D) are consistent with any of the three recommended strategies? (Hint: Standard Deviation = 117.6) (Circle ALL that apply.)

Reported cases of desease A per 100,000 population by census tract, City of Dixon, 1991

Census Tract	Cases per 100,000 population
1	170.5
2	0.0
3	70.0
4	40.0
5	115.0
6	42.1
7	453.5
8	0.0
9	35.1
10	50.3
11	0.0
12	0.0
13	186.4
14	49.9
15	48.9
Total	**1,262.2**

A. 0.0
0.1- 84.1
84.2-201.7
201.8-453.5

B. 0.0- 35.1
35.2- 50.3
50.4-453.5

C. 0.0- 50.0
0.1-100.0
100.1-200.0
200.1-453.5

D. 0.0-113.4
113.5-226.8
226.9-340.2
340.3-453.6

8. The *main distinction* between an arithmetic-scale line graph and a semilogarithmic-scale line graph is that the arithmetic scale:

   A. measures the rate of change between successive points on a graph

   B. is preferred when the range of values to be graphed is very large

   C. uses equal distances on each axis to represent equal quantities

   D. is the best method of showing changes in the magnitude of numbers

9. Which type of graph is recommended for showing annual mortality rates for Disease Z, for 1940 to 1990? (Circle ALL that apply.)

   A. Arithmetic-scale line graph

   B. Semilogarithmic-scale line graph

   C. Histogram

   D. Frequency polygon

10. Which of the following sets of values would be *inappropriate* for identifying equidistant intervals on the *y*-axis of a semilogarithmic-scale line graph?

    A. 1, 10, 100, 1,000

    B. 10, 20, 30, 40

    C. 7, 70, 700, 7,000

    D. 0.003, 0.03, 0.3, 3

11. Bar charts may be distinguished from histograms at a glance because:

    A. bar charts are not used for time series data

    B. histograms are used to display discrete data

    C. bar charts are based on area under the curve

    D. histograms do not have spaces between consecutive columns

12. Which of the following statements are true of an epidemic curve? (Circle ALL that apply.)

    A. An epidemic curve is a histogram.

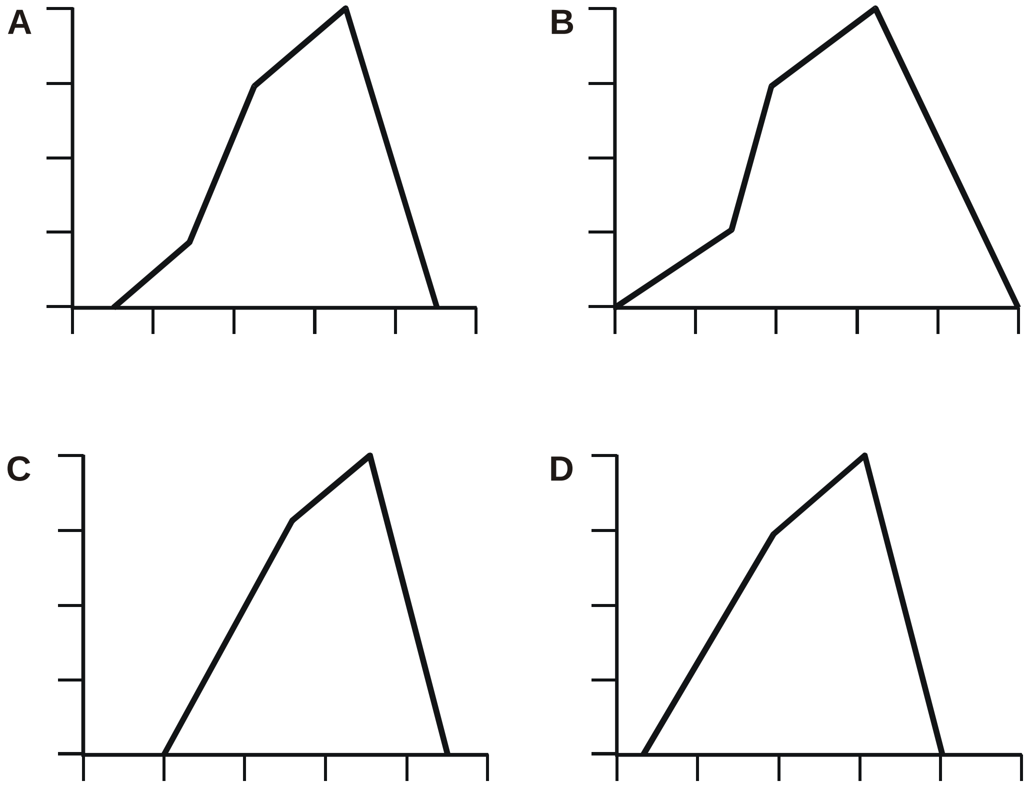
    B. An epidemic curve shows number of cases by date of exposure.

    C. An epidemic curve should begin with the first case of the outbreak.

    D. An epidemic curve should use time intervals on the *x*-axis of about 1/2 of the incubation period.

13. Which one of the following methods of closing a frequency polygon on the horizontal axis is correct?

Figure 4.35
Correct and incorrect methods of closing a frequency polygon



14. Which type of graph or chart would be appropriate for graphing deaths over time for a cohort of 100 alumni from the Class of 1907? (Circle ALL that apply.)

A. Bar chart

B. Cumulative frequency curve

C. Histogram

D. Survival curve

Choices for questions 15-20:

A. arithmetic-scale line graph

B. bar chart

C. series of box plots

D. series of dot plots

E. frequency polygon

F. scatter diagram

15. Number of cases by a continuous variable ________

16. Number of cases by a discrete (noncontinuous) variable ________

17. Mean value of one continuous variable by a second continuous variable ________

18. Median value of continuous variable by a discrete (noncontinuous) variable ________

19. Each value of one continuous variable by a second continuous variable ________

20. Each value of a continuous variable by a discrete (noncontinuous) variable ________

21. What type of graph is most appropriate for comparing rates of change of disease occurrence over several years?

A. Arithmetic-scale line graph

B. Semilogarithmic-scale line graph

C. Histogram

D. Frequency polygon

22. What type of graph is most appropriate for comparing the magnitude of events which have occurred in different places, but no map is available?

A. Arithmetic-scale line graph

B. Bar chart

C. Frequency polygon

D. Histogram

23. Which type of chart could be used to display the relative size of different causes of death by sex? (Circle ALL that apply.)

    A. One simple bar chart

    B. One grouped bar chart

    C. One stacked bar chart

    D. 100% component bar chart (multiple bars)

    E. One pie chart

24. The best choice for displaying years of potential life lost for different causes of death is:

    A. one simple bar chart

    B. one grouped bar chart

    C. one stacked bar chart

    D. 100% component bar chart (multiple bars)

25. Which of the following statements are true concerning an area map compared with a spot map? (Circle ALL that apply)

    A. The area map shows the location of a case or event more specifically.

    B. Only the area map can portray risk or rate of disease.

    C. Only the area map can portray two or more cases at the same location.

    D. An area map can portray *rates*, but only a spot map can show *numbers* of cases.

Answers are in Appendix J
If you answer at least 20 questions correctly, you understand
Lesson 4 well enough to go to Lesson 5.

References

1. Alter MJ, Ahtone J, Weisfuse I, Starko K, Vacalis TD, Maynard JE. Hepatitis B virus transmission between heterosexuals. JAMA 1986; 256:1307-1310.
2. Centers for Disease Control. Chronic Disease Supplement, 1987. Deaths from cervical cancer—U.S., 1984-1986. MMWR 1989;38:38.
3. Centers for Disease Control. HIV/AIDS Surveillance Report. November 1990.
4. Centers for Disease Control. Manual of reporting procedures for national morbidity reporting and public health surveillance activities. July 1985.
5. Centers for Disease Control. Progress toward eradicating poliomyelitis from the Americas. MMWR 1989;39:33.
6. Centers for Disease Control. Infant mortality among racial/ethnic minority groups, 1983-1984. MMWR 1990;39:SS-3.
7. Centers for Disease Control. St. Louis encephalitis — Florida and Texas, 1990. MMWR 1990;39:42.
8. Centers for Disease Control. MMWR 1991;40:4.
9. Centers for Disease Control. Nutritional assessment of children in drought-affected areas — Haiti, 1990. MMWR 1991;40:13.
10. Centers for Disease Control. Cigarette smoking among adults — United States, 1988. MMWR 1988;40:44.
11. Centers for Disease Control. National Institute of Occupational Safety and Health. National Traumatic Occupational Fatalities Database.
12. Centers for Disease Control. Summary of notifiable diseases, United States, 1989. MMWR 1989;38(54).
13. Centers for Disease Control. Health status of Vietnam veterans. Volume 3: Medical Examination. 1989.
14. Creech JW. Effective oral presentations. Epi in Action Course, Centers for Disease Control, 1988.
15. Dicker RC, Webster LA, Layde PM, Wingo PA, Ory HW. Oral contraceptive use and the risk of ovarian cancer: The Centers for Disease Control Cancer and Steroid Hormone Study. JAMA 1983;249:1596-1599.
16. Fingerhut MA, et al. Cancer mortality in workers exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin. New Engl J of Med 1991; 324:212-218.
17. Hadler SC, et al. Occupational risk of hepatitis B infection in hospital workers. Infect Ctrl 1985; 6:24-31.
18. Kleinman JC, Donahue RP, Harris MI, Finucane FF, Madans JH, Brock DB. Mortality among diabetics in a national sample. Am J Epidemiol 1988;128:389-401.

19. Lettau LA, et al. Outbreak of severe hepatitis due to delta and hepatitis F viruses in parenteral drug abusers and their contacts. New Engl J of Med 1987; 317:1256-1262.

20. McKenna M, Wolfson S, Kuller L. The ratio of ankle and arm arterial pressure as an independent predictor of mortality. Athero 1991; 87:119-128.

21. National Center for Health Statistics. Advance report of final mortality statistics, 1987. Monthly vital statistics report; vol 38, no 5 supp. Hyattsville, MD: Public Health Service. 1989.

22. Schoenbaum SC, Baker O, Jezek Z. Common source epidemic of hepatitis due to glazed and iced pastries. Am J Epidemiol 1976;104:74-80.

23. Schreeder MT, et al. Hepatitis B in homosexual men: prevalence of infection and factors related to transmission. J Infect Dis 1982; 146:1.

24. Sutter RW, Patriarca PA, Brogran S et al. Outbreak of paralytic poliomyelitis in Oman. Evidence for widespread transmission among fully vaccinated childern. Lancet 1991; 338:715-20.

25. Tufte ER. The visual display of quantitative information. Cheshire, CT: Graphics Press, 1983.

26. Wells DL, Hopfensperger DJ, Arden NH, et al. Swine influenza virus infections. JAMA 1991; 265:478-481.

27. Williamson DF, Parker RA, Kendrick JS. The box plot: A simple visual method to interpret data. Ann Intern Med 1989; 110:916-921.